History's Most Revolutionary Innovation: AI and Its Origins and Impact

by

Victor Menaldo

University of Washington

Publisher: Cambridge University Press & Assessment

Dedication

this is for the robots who may inherit the earth one day; I hope you've learned enough from our mistakes to not repeat them.

TABLE OF CONTENTS

Note: Page numbers will be generated after copyediting/typesetting.

Chapter 1. A Fourth Industrial Revolution?

Chapter 2. What's an Industrial Revolution?

Chapter 3. The Creative Destruction Paradigm

Chapter 4. Superstar Dynamics

Chapter 5. High Tech Superstar, Industry Tech Leaders

Chapter 6. Smartphone Supply Chain & Interoperability

Chapter 7. AI's Antitrust Seedbed

Chapter 8. Digital Platforms, Data Flywheels

Chapter 9. Industrial Organization of the AI Stack

Chapter 10. AI Standardization & Organizational Adaptation

Chapter 11. Macroeconomics of AI

Chapter 12. The Fragmented Revolution

LIST OF ABBREVIATIONS

Abbreviation	Meaning
4G	Fourth-generation mobile standard
5G	Fifth-generation mobile standard
AC	Alternating Current
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AMPS	Advanced Mobile Phone System
API	Application Programming Interface
ASIC	Application-Specific Integrated Circuit
AWS	Amazon Web Services
BLS	Bureau of Labor Statistics
CAFC	Court of Appeals for the Federal Circuit
CAPEX	Capital Expenditures
CBA	Cost–Benefit Analysis
CDMA	Code Division Multiple Access
CDP	Creative Destruction Paradigm
CEA	Council of Economic Advisers
CI/CD	Continuous Integration / Continuous Deployment
CPI	Consumer Price Index
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DMCA	Digital Millennium Copyright Act
DNS	Domain Name System
DoJ	U.S. Department of Justice

DSP	Digital Signal Processor
DUV	Deep Ultraviolet (lithography)
E-SIGN	Electronic Signatures in Global and National Commerce Act
EDA	Electronic Design Automation
EO	Executive Order
EPC	Evolved Packet Core
EU	European Union
EUV	Extreme Ultraviolet (lithography)
FCC	Federal Communications Commission
FLOP/s	Floating-Point Operations per Second
FLOPs	Floating-Point Operations (count)
FP64	64-bit (double-precision) Floating Point
FRAND	Fair, Reasonable, and Non-Discriminatory (licensing terms)
FTC	Federal Trade Commission
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation (EU)
GPT	General Purpose Technology; also used in model names as Generative Pre-trained Transformer (e.g., GPT-3)
GRPO	Group Relative Policy Optimization
GSM	Global System for Mobile Communications
HHI	Herfindahl–Hirschman Index
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IETF	Internet Engineering Task Force

IMF	International Monetary Fund
IoT	Internet of Things
IP	Intellectual Property
IPR	Intellectual Property Rights
ISDN	Integrated Services Digital Network
KV cache	Key–Value cache (for attention during inference)
LLM	Large Language Model
LTE	Long-Term Evolution (4G mobile standard)
MIMO	Multiple-Input Multiple-Output
MIPS	Million Instructions Per Second
MIT	Massachusetts Institute of Technology
MLOps	Machine Learning Operations
MME	Mobility Management Entity
MoE	Mixture of Experts
NAFTA	North American Free Trade Agreement
NBER	National Bureau of Economic Research
NEC	National Economic Council
NIE	New Institutional Economics
NIST	National Institute of Standards and Technology
NPU	Neural Processing Unit
NVLink	NVIDIA high-speed GPU interconnect
NVMe	Non-Volatile Memory express
OECD	Organisation for Economic Co-operation and Development

OFDMA	Orthogonal Frequency-Division Multiple Access
OIRA	Office of Information and Regulatory Affairs
OMB	Office of Management and Budget
OS	Operating System
OSAT	Outsourced Semiconductor Assembly and Test
PC	Personal Computer
PCC	Private Cloud Compute (Apple)
R&D	Research and Development
RAG	Retrieval-Augmented Generation
RDU	Reconfigurable Dataflow Unit
RIA	Regulatory Impact Analysis
RISC	Reduced Instruction Set Computing
RLHF	Reinforcement Learning from Human Feedback
S3	Amazon Simple Storage Service
SC-FDMA	Single-Carrier Frequency-Division Multiple Access
SEP	Standard-Essential Patent
SIM	Subscriber Identity Module
SLM	Small Language Model
SSNIP	Small but Significant and Non-transitory Increase in Price
SSO	Standards-Setting Organization
TCP/IP	Transmission Control Protocol / Internet Protocol

TFP	Total Factor Productivity
TPU	Tensor Processing Unit
TRIPS	Trade-Related Aspects of Intellectual Property Rights
TSMC	Taiwan Semiconductor Manufacturing Company
UETA	Uniform Electronic Transactions Act
URL	Uniform Resource Locator
USPTO	United States Patent and Trademark Office
USTR	United States Trade Representative
VoLTE	Voice over LTE
W3C	World Wide Web Consortium
WTO	World Trade Organization
x86	Intel instruction set architecture family

LISTS (FIGURES / TABLES)

LIST OF FIGURES

Figure 2.1. Moore's Law Over Time.

Figure 2.2. Fewer Ideas as a Market Failure.

Figure 2.3. Different Types of Goods

Figure 4.1. Foreign-value-added in US Manufacturing Exports, % Manufacturing Exports

Figure 4.2. High-Value Business Services Exports, % of Total Services.

Figure 4.3. The Evolution of US Economy Towards Intangible Capital.

Figure 4.4. US Net IP Payment Receipts, % GDP.

Figure 5.1. : Power Law Fitted to Patent Forward Citation Data, High Tech Firms.

Figure 5.2. : Visualizing the Semiconductor Network.

Figure 5.3. : Power Law Fitted to Probability Degree Distribution

Figure 5.4. : Vertical Network and the Role of Texas Instruments

Figure 5.5. : Vertical Network and the Role of Intel

Figure 5.6. : Vertical Network and the Role of Qualcomm

Figure 6.1. : The Costs of Running a Chip Foundry—TSMC's Escalating Expenses.

Figure 9.1. The AI Supply Chain.

Figure 9.2. AI Chip Performance, 2017-2025.

Figure 9.3. Quasi Vertically Integrated AI Stack.

Figure 10.1. The Epic AI Infrastructure Construction Boom.

Figure 10.2. Hyperscaler CapEx Spending, 2019-2025.

Figure 10.3. Exponential Adoption Pattern of Web Visitors to ChatGPT Chatbot.

Figure 11.1. A Partial-equilibrium Labor-market Model for a Single Sector.

LIST OF TABLES

Table 3.1. Comparing the CDP to Neo-liberalism and the Washington Consensus

Table 3.2. Key CDP Changes to IP

Table 9.1. Strategic Approaches to AI Stack Integration

Table 11.1. Summarizing the Theoretical Framework's Key Insights

Table 11.2. The Vampire versus the Superstar

PREFACE

I came to this project as a naïve lamb heading for slaughter, but with eyes wide open because I was not an expert by any stretch of the term. I got interested in AI around the time my second son, Thomas, was born, because I was keeping up with the news coming out of the computer world and it seemed like things were changing in subtle but profound ways. Around then, 2016, was the year that AlphaGo beat Lee Sedol, Google Translate quietly flipped over to neural machine translation, and deep learning started leaking out of the lab and into everyday products in a way you could actually feel. It would of course take a few more years for ChatGPT to burst on the stage and really change everything, but as early as 2016 it already seemed like something important was brewing, so I began to take note and let everybody know that I had made up my mind and was writing a book about AI. Little did I know how many questions I'd get once chatbots changed everything.

What I did at first was to plug-and-play with my frameworks and tools from political science, economics and history that I'd honed studying things like the political economy of natural resources, industrialization, and finance, regime change dynamics, and the economic development process in poorer countries. I realized that this could be an asset as the world around me turned in a less predictable, more personalistic, even crueler direction, and as I began to ask questions about AI that I could not find satisfactory answers for in the literature. Any literature.

I therefore turned to history, industrial organization theory, basic microeconomic and macroeconomic principles and started reading everything about AI I could get my hands on. This was all very analog at first, relying on stacks of printed paper and folders and physical Economist magazine articles as far as the eye can see. Once AI had caught up to my needs, it accelerated my progress, as I experimented with ChatGPT, Claude, Perplexity and Gemini to increase my productivity by using AI as a souped-up search engine, tireless RA, a sounding board, ersatz therapist, an unsparing critic—I'm looking at you ChatGPT 5.2—and, more practically, as a meta spreadsheet containing all my knowledge on the topics in this book (and a lot of its own). It also became a pretty cool word processor, and a statistician like none I'd ever met before. What I never let it do was write for me, however, because I realize that there are just some things that humans still have both an absolute and comparative advantage in, and logical thought and the ability to reason through problems based on first principles is one of them!

I have learned more over the last 10 years than I ever have before in my life. The AI revolution was not only fascinating to read about and write about and try to figure out but it was also fun to experience in real time as I grappled with huge questions about where innovation comes from, why it evolves the way it does, when its benefits pay off, if at all,

and who enjoys the lion's share of those. More personally, I have a lot at stake in these issues because it's my future too, and especially my children's.

I can't say I have satisfactory answers for any of the questions I set out to answer yet; but I do consider this a first, publishable attempt to formalize an approach that I hope readers will find useful and build on…

ACKNOWLEDGMENTS

I'd like to thank Will Ambrusio for hosting me at Holden Village, nestled in the incomparable North Cascades, in the summer of 2025 and for the opportunity to teach on many of the themes that occupy the core of this book, including industrial revolutions and artificial intelligence. The conversations and feedback from Holden's attendees were generous, wide-ranging, and deeply formative. I am also grateful to Amanda Clayton for hosting me at UC Berkeley for the Comparative Politics seminar in March 2025, where I presented Chapter 12, and to Beatrice Magistro for inviting me to Caltech for a conference on the political economy of AI in September 2024, where I presented Chapter 7.

I thank Jake Grumbach for including me in a small electronic seminar on American political economy and innovation in Spring 2021, where I presented a very primitive version of Chapter 5, and I thank the Political Economy Forum at the University of Washington for opportunities to present multiple chapters on two occasions.

I am especially grateful to the Hoover Institution IP2 Summer Teaching Institute for inviting me to serve as core faculty from 2014 to 2018. I learned an enormous amount from colleagues there—especially Steve Haber, the late Alexander Galetovic, Naomi Lamoreaux, Zorina Khan, Richard Epstein, Scott Kief, Jonathan Barnett, Wess Hartman, Ross Levine, Latika Chaudhary, and Josh Wright—about all manner of useful things. Many of them appear in these pages, in fact. These amazing researchers and teachers sharpened my thinking about what evidence-based policymaking really means and taught me to be skeptical about regulation for regulation's sake.

For invaluable comments made at Political Economy Forum lunches at UW, I thank James Long, Chris Adolph, John Wilkerson, Mark Smith, Gabriella Levy, Asli Cansunar, Susan Whiting, Ramses Llobet, and Brian Leung. For comments across these outside talks, I thank Ethan Bueno de Mesquita, Beatrice Magistro, and Mike Alvarez. Finally, for superb research assistance, I thank Lucas Owen, Nicolas Wittstock, Dan Dembowy, Anthony Wang, and Elizabeth Martyross.

Finally, my sincerest gratitude goes to my CUP editor Robert Dreesen, who made this an extremely pleasant and easygoing experience. Thanks for your steady guidance and support.

CHAPTER 1. A FOURTH INDUSTRIAL REVOLUTION?

The standardization, commercialization, and mass diffusion of transformative technologies such as steam engines, electrical motors, and microchips has helped humans produce almost everything they make at much greater scale and at a much lower cost—ranging from textiles to foodstuffs to machines. In turn, this phenomenon fueled so-called Hockey Stick-shaped economic progress. By bending the development curve, industrialized economies escaped the "Malthusian Trap"—an ecological tragedy whereby exponential population growth consistently outpaces merely linear improvements in agricultural production, resulting in recurring famines, disease, and mass death (McCloskey 2016).

Especially between 1920 and 1970, rapid productivity growth followed widespread electrification, underwriting substantial improvements in real wages and bankrolling generous welfare states. Affordable electric appliances such as refrigerators, washing machines, and vacuums became ubiquitous, freeing people from tedious, if not backbreaking, housework and similar chores. Leisure time went from a luxury for the few to a commodity enjoyed by the many (Gordon 2016).

Fast forward to 2025, when another technology not so far removed from electricity—artificial intelligence—was making a huge splash. Chatbots excelled at drafting documents and presentations, streamlining research, generating images, and writing and debugging code. Despite flaws that this book will explore, generative AI increasingly functioned as a universal interface: between people and machines (as in software engineering), between machines and machines (as in factory robotics), and between computers and computers (as in agentic workflows). A user could type, “build a booking tool that sends text message reminders,” and the system could generate a working prototype—designing an interface, writing core logic, and

automating much of the invisible plumbing that once required a professional engineering team. The machinery of deployment did not disappear, but it became increasingly abstracted away from the user.

What's more, AI began to automate parts of the automation process itself. Models did not merely predict the next word or pixel; they increasingly helped generate code, design experiments, and optimize systems. In biology, systems like AlphaFold accelerated discovery by extracting structure from vast scientific data and enabling new hypotheses and experiments (Jumper et al. 2021). In software engineering, AI-assisted coding systems began to generate working code that could be used to build and refine new tools (Chen et al. 2021). And in operations, AI was increasingly used to optimize complex logistical systems, including supply chains, for cost, speed, and resilience (Ivanov 2021).

This type of recursive self-improvement recalled I. J. Good's famous declaration that "the first ultraintelligent machine is the last invention that man need ever make" (Good 1965: 33). Readers of the distant future, or even immediate future, may laugh at the wonder this tries to convey, but for the denizens of the mid 2020s, AI indeed felt like history's most important invention.

The Book's Thesis

This book argues that to best understand AI as a potentially transformative technology, we should heed the lessons of steam engines, electricity, and microchips—the General Purpose Technologies that powered the first three industrial revolutions. Like its predecessors, AI exhibits exponential performance improvements and S-curve adoption patterns. But history teaches that technical breakthroughs alone do not rewire economies. To get promising technologies such as driverless cars on every curbside, it will take a massive sorting of liability

rules, safety regulations, traffic laws, and privacy and cybersecurity protocols. Transformative technologies require decades of such institutional scaffolding—property rights, liability rules, technical standards, and organizational adaptation—before their productivity potential is realized.

The book's central claim is that AI's rapid scaling was not an accident of technology alone. It was the product of a historically contingent political, legal, and regulatory landscape—what I call the Creative Destruction Paradigm—that emerged from a bipartisan transformation of the American political economy between the late 1970s and early 2000s. Through reforms to intellectual property, telecommunications, antitrust, electronic contracting, and data governance, the CDP reduced transaction costs in ways that enabled digital platforms to amass the massive datasets required to train Large Language Models. Understanding how that landscape emerged—and how it is now fragmenting—is essential for anyone grappling with AI's implications.

Several additional arguments structure the book. Taken together, they trace a single arc: the institutional and market reforms that enabled intangible-capital specialization and vertically disintegrated supply chains also produced an innovation commons whose emblematic product was the smartphone—an ecosystem that generated the data and platform architectures that, in turn, made today’s AI possible, even as that enabling environment began to fragment.

First, I show that the American economy's shift toward intangible capital—R&D, IP, software, and algorithms—was enabled by the vertical disintegration of global supply chains. As US firms specialized in high-value design while outsourcing physical production, a handful of "superstar" firms captured outsized returns.

Second, using novel patent citation network analysis, I demonstrate that these technological leaders simultaneously dominated their sectors and diffused innovations across supply chains, creating an innovation commons that accelerated industrywide progress.

Third, I explain how the smartphone supply chain became the *éminence grise* of this globally distributed innovation system; international standards-setting organizations and FRAND commitments helped mold proprietary patents into shared infrastructure by helping ensure that essential technologies are licensed on fair, reasonable, and non-discriminatory terms.

Fourth, I explore how digital platforms that generated geo-coded, dynamic behavioral data evolved to do this so well, achieved a global scale, and how this helped power today's AI.

Fifth, I develop a macroeconomic model that formalizes the conditions under which an AI productivity boom translates into rising living standards versus stagnant wages. The model isolates the role of standardization, organizational adaptation, and markup dynamics in determining whether efficiency gains flow to workers or are captured by "gatekeepers."

Sixth, I document how, by 2025, the world economy in general and the AI industry in particular fragmented into competing governance regimes. In the United States circa 2025 politics had seemed to devolve into a patrimonial hybrid where market mechanisms persist, but outcomes increasingly depend on political relationships. China had consolidated state-centric managerialism, requiring ideological alignment as a condition of market entry. The European Union pursued a regulatory-first approach, betting that a heavier hand might endow it with a competitive advantage in model alignment and AI safety. This fragmentation marked the end of the borderless digital order that characterized the rise of the internet, with profound implications for the trajectory of AI innovation.

UNDERSTANDING AI: A PRIMER FOR POLITICAL ECONOMISTS

In what follows I briefly explain what AI is and isn't; its purpose is to supply technical building blocks and a shared vocabulary for the institutional and economic analysis that follows. Readers already familiar with the basics can skim this primer and return as needed or even skip it altogether. For those of you sticking around, buckle up.[1]

From Rules to Learning

Traditional software operates by following explicit instructions written by programmers: if a customer's balance falls below zero, charge a fee; if a password matches, grant access. These rule-based systems excel at well-defined tasks but struggle with ambiguity. How would one write rules to recognize a cat in a photograph, or to determine whether a movie review is positive or negative?

Machine learning inverts this logic. Rather than programmers specifying rules, machine learning systems learn statistical patterns from data and adapt their outputs accordingly (Russell and Norvig 2021). In supervised learning, the dominant approach, each input is paired with a known correct output—thousands of images labeled "cat" or "not cat," millions of sentences tagged as "positive" or "negative." By iteratively adjusting internal parameters to minimize prediction errors, these systems eventually learn to classify new, unseen inputs with remarkable accuracy.

This shift from explicit programming to learned patterns explains why machine learning underpinned most AI advances during the early 2020s. Instead of anticipating every scenario in advance, engineers could train systems on massive datasets and let the algorithms discover the

[1] In an online supplementary appendix, I do a much deeper dive into the technical details, as well as the history of AI and its more recent evolution (see Menaldo 2026: Section 3).

relevant features themselves. Hardcoding rules and a rote set of potential responses became passe and the stochastic discovery of new patterns surprising to both researchers and machines was in vogue.

Neural Networks: The Architecture of Modern AI

Neural networks represent machine learning's most transformative development. Inspired loosely by biological neurons, these systems consist of interconnected processing units arranged in layers. Data enters through an input layer, flows through one or more "hidden" layers where patterns are detected, and emerges from an output layer as a prediction or classification (LeCun et al. 2015; Goodfellow et al. 2016).

The key insight is hierarchical abstraction. In image recognition, early layers detect simple features like edges and corners; middle layers combine these into shapes and textures; deep layers recognize complex objects like faces or vehicles. Each layer builds upon the previous one, enabling the network to learn increasingly sophisticated representations from raw data.

Deep learning refers to neural networks with many hidden layers—sometimes dozens or hundreds—that can capture extraordinarily complex patterns. This depth, combined with massive datasets and specialized hardware, enabled breakthroughs in computer vision, speech recognition, and natural language processing that had eluded researchers for decades.

How Large Language Models Generate Text

The AI systems that captured public attention in the early 2020s—ChatGPT, Claude, Gemini—are LLMs built on an architecture called the "transformer" (Vaswani et al. 2017). Understanding how they work illuminates both their remarkable capabilities and their fundamental limitations.

An LLM operates as a sophisticated prediction engine. Given a sequence of words, it predicts the most likely next word. When a user types "The weather is," the model might assign a

35 percent probability to "nice," 25 percent to "beautiful," and 15 percent to "cold." It selects a word based on these probabilities, appends it to the input, and repeats—generating text one word at a time.

The transformer architecture's key innovation is "self-attention," a mechanism that allows the model to weigh the relationships between all words in a passage simultaneously. When processing "The bank by the river was steep," self-attention helps the model recognize that "bank" relates to "river" and "steep," indicating a riverbank rather than a financial institution. This capacity to consider context across long passages enables coherent, contextually appropriate responses.

Critically, these capabilities emerge from scale. LLMs are trained on hundreds of billions of words scraped from the internet—books, articles, social media posts, code repositories, and more. Through this exposure, they internalize statistical regularities about language: grammar, facts, reasoning patterns, and stylistic conventions. The "understanding" they exhibit is an emergent property of identifying complex relationships between words across vast training data—not comprehension in the human sense.

The Compute Explosion

The economic and political significance of modern AI cannot be understood without grasping the extraordinary resources required to build frontier systems. Training a large language model is not a software project in the traditional sense; it is an industrial undertaking requiring purpose-built infrastructure. Training is the up-front process of fitting a model’s parameters using massive compute; inference is the ongoing process of running the trained model to answer user queries—and at scale, inference can rival or exceed training in total cost.

The computational resources required to train prominent AI systems expanded exponentially, with training compute doubling approximately every six months after 2010 (Sevilla et al. 2022). What began as modest academic experiments grew to require billions of calculations for systems like AlexNet in 2012, then trillions for GPT-2, and eventually numbers so large they defy intuition for GPT-4 and its successors.

This exponential scaling fundamentally altered who could participate at the frontier. When state-of-the-art systems required modest compute, university laboratories could compete with industry. Once competitive training runs demanded hundreds of millions of dollars in specialized hardware, electricity, and cooling infrastructure, the center of gravity shifted decisively to well-capitalized corporate labs (Ahmed and Wahed 2020). By 2025, only a handful of organizations—OpenAI, Anthropic, Google DeepMind, Meta AI, and a few others—possessed the resources to train frontier models.

The capital intensity also explains the talent wars splashed against the pages of industry press during 2025. Nine-figure compensation packages for elite researchers existed precisely because the researchers who know how to exploit billion-dollar compute clusters were extraordinarily scarce. The constraint had shifted from ideas to execution—data, compute, and the people who could transform both into working models.

A Brief History: From Dartmouth to ChatGPT

The field of artificial intelligence was formally established at the 1956 Dartmouth Conference, where researchers including John McCarthy and Marvin Minsky coined the term and predicted rapid progress toward human-level machine intelligence (McCarthy et al. 1956). Early optimism soon collided with technical limitations, triggering periods of reduced funding and interest known as "AI winters."

The contemporary deep learning revolution traces to breakthroughs in the 1980s and 2000s, when researchers developed techniques for training neural networks with many layers. The watershed moment came in 2012, when a deep neural network called AlexNet dramatically outperformed traditional approaches on an image recognition benchmark, demonstrating that deep learning could solve problems that had stymied the field for decades (Krizhevsky et al. 2012).

The transformer architecture, introduced by Google researchers in 2017, removed key bottlenecks that had limited earlier approaches to language (Vaswani et al. 2017). OpenAI's GPT series demonstrated that scaling transformers to enormous sizes produced surprising capabilities—including the conversational fluency that made ChatGPT a global phenomenon upon its November 2022 release (Brown et al. 2020).

By 2024-25, the frontier had pivoted toward "reasoning models" that could deliberate on problems rather than generating immediate responses. Techniques like reinforcement learning from human feedback (RLHF) aligned model outputs with human preferences, while chain-of-thought prompting encouraged step-by-step reasoning (Christiano et al. 2017; Ouyang et al. 2022). These advances improved performance on complex tasks but did not resolve the fundamental limitations discussed below.

The Limits of Pattern Matching

For all their impressive capabilities, contemporary AI systems exhibit a crucial limitation that shapes their economic potential: they learn statistical associations, not causal understanding (Marcus 2025a; 2025b; Chollet 2019).

When an LLM answers a medical question correctly, it does so by pattern-matching against similar text in its training data—not by understanding anatomy or physiology. This

distinction matters because pattern-matching fails when problems differ from training examples in subtle ways. Studies have shown that reasoning models collapse on simple puzzles when complexity increases beyond certain thresholds, and that models often cannot reliably follow explicit algorithms even when provided in the prompt (Shojaee et al. 2025).

This brittleness manifests in practical deployments. AI systems "hallucinate" plausible-sounding but fabricated information. They exhibit inconsistent performance across contexts. Security researchers have demonstrated that carefully crafted inputs can cause even hardened systems to fail 50 to 90 percent of the time (Huang et al. 2025). A 2025 study found that 95 percent of organizations piloting generative AI reported zero return on their efforts after six months (Challapally et al. 2025).

While these limitations do not negate AI's transformative potential—electricity and semiconductors also faced decades of practical obstacles before reaching their full impact, they do suggest that the path from laboratory breakthrough to economywide productivity gains will be neither automatic nor immediate. As subsequent chapters will show, realizing AI's potential requires institutional frameworks, organizational adaptation, and regulatory structures that take time to develop.

Evidence of an Industrial Revolution

The first three industrial revolutions transformed the world. The first cemented the British Empire and consolidated the scientific and liberal revolutions; it also saw increased globalization in terms of greater trade volumes and migration (O'Rourke and Williamson 1999). The second helped spread industrialization and mass consumption (Landes 2008; Hounshell 1984). The third was coeval with another epic wave of globalization that saw democracy and markets spread

around the globe, especially after the end of the Cold War. The fourth should be no different, heralding massive economic, social, and political changes on multiple fronts.

But how do we know that the early 2020s AI boom was indeed an industrial revolution? How does one know if one is living through an industrial revolution without the benefit of hindsight? Fortunately, there are several ways to explore whether another industrial revolution was upon us.

Intensive Capital Buildouts

One possible way to know if an industrial revolution is happening is to visualize how the infrastructure associated with its signal technologies swells during its nascent, "installation" phase (Perez 2002). During the original industrial revolution, a vast network of coal mines, railways, ports, bridges, tunnels, canals, and turnpike roads rapidly swelled over the course of just a few decades. During the second industrial revolution, starting around 1870, innovations in steelmaking, electrification, and fossil fuels induced massive investments in electrical generation plants and transmission lines, oil rigs, pipelines, petroleum refineries, vast railroad networks, and football-sized manufacturing facilities and warehouses. During the third industrial revolution, huge mainframes, server farms, fiber optic cables, and cell towers sprang up throughout the world to create a global digital communications network.

The AI Revolution's heavy, visceral footprint is vividly exemplified by the rapid proliferation of massive data centers in communities across central Washington state, the gateway to the Columbia River's abundant hydroelectric power and its cooling waters. Constructed with billions in investment, these facilities mirror how coal mines and railway hubs sprang up across the American frontier during its nineteenth-century industrialization. Building them involves leveling vast rural plots, installing reinforced concrete shells, and integrating

redundant power systems tied directly to the Columbia River's dams for sustainable, low-cost energy. Inside, they house endless rows of server racks, Graphics Processing Units (GPUs), and cooling arrays, humming with the computational might to train AI models on petabytes of data.

By the early 2020s, places like Quincy, a small, bucolic central Washington town dotted with potato farms, hosted almost 50 massive data centers, including Microsoft's Columbia Data Center, which sprawled across 800,000 square feet on 270 acres, with a power draw exceeding 50 MW—enough to light up a small city. This transformed it from a sleepy outpost that generated modest revenues tied to farming into a thriving community plugged into the AI revolution that boasted high-paying jobs and received $54 million in taxes in 2025, which it used to fund state-of-the-art schools, community centers, and roads (Nickelsburg et al. 2025).

Exponential Adoption Patterns

Another surefire sign that an industrial revolution is underway is to witness exponential adoption patterns. Consider driverless cars. As early as 2005, Stanford students and professors working out of a then-unheralded AI laboratory built a robot named Stanley that won the DARPA Grand Challenge for driverless cars. Members of that team were subsequently recruited to lead Google's secret self-driving project, later branded Waymo, which the company officially revealed in 2010 (Cohen 2025). Commercialization efforts followed ploddingly. Self-driving programs run by Tesla and Waymo poured billions of dollars into bringing driverless cars to market (Kolodny and Elias 2024). Waymo's team completed the world's first completely autonomous ride on a public road in 2015; it was only by 2020 that its driverless car service officially opened for business.

Like what happened with ChatGPT, a classic exponential takeoff soon took hold. While it took until late 2023 for Waymo to reach one million cumulative paid trips, that number quintupled to five million by the end of 2024. In the first few months of 2024, Waymo was

completing only tens of thousands of trips per month. By mid-year, volume began to double roughly every quarter, surging from approximately 50,000 trips in April to 100,000 in July, before accelerating to 230,000 in September and ending the year at nearly 440,000 trips in December. The momentum intensified in early 2025, with monthly ride counts climbing from roughly 490,000 in January to 700,000 by June, pushing the cumulative total past 10 million in the first half of the year alone. Continuing this eye-popping trajectory, Waymo exceeded 20 million fully autonomous trips by the end of 2025 (Waymo 2025).

Beyond autonomous vehicles, several other AI applications exhibited the classic S-curve signature of industrial diffusion: slow early experimentation followed by rapid adoption once reliability and integration costs fell. Coding copilots scaled from niche developer tools into mainstream productivity infrastructure. Customer-support automation diffused rapidly. In media, generative image and video tools spread with viral speed. For enterprise uses, retrieval-augmented search and document summarization similarly accelerated.

Exponential Performance Improvements

Yet another hallmark of industrial revolutions is exponential improvement in technological performance as investment in capital and R&D intensifies (Galetovic 2021). If we measure AI progress using METR's "50%-task-completion time horizon"—the length (for humans) of tasks that frontier agents can complete autonomously with 50% success—the frontier has advanced at a striking pace, with the time horizon doubling roughly every seven months since 2019 (Kwa et al. 2025). On this metric, the frontier moved from seconds-long tasks in 2019 to around an hour by early 2025 for models such as Claude 3.7 Sonnet and contemporaneous reasoning-oriented systems. This rate of capability expansion is far faster than the historical cadence associated with Moore's Law, where transistor counts doubled roughly every 18–24 months.

Multiple Applications

Industrial revolutions are also characterized by the simultaneous flourishing of multiple practical uses—and the AI revolution is no different. AI applications that employed computer vision, reasoned probabilistically, and performed search, optimization, and planning were able to harness these various capabilities to learn, reason, perceive, communicate, and make decisions at or above human levels across several domains.

In healthcare, providers employed two distinct types of AI. They used transformer-based language models fine-tuned on medical literature and patient records to assist with diagnosis research and treatment planning (Miotto et al. 2016). Separately, they deployed specialized computer vision models, trained on medical imaging datasets, to analyze X-rays, MRIs, and other medical imaging data (Litjens et al. 2017). The larger implication was a significant leap towards a new standard of care, where a physician's expertise was augmented by the speed and pattern-recognition abilities of AI.

In industrial settings, the Internet of Things—a network of interconnected machinery and sensors collecting massive amounts of data—powered AI applications. Manufacturers deployed computer vision models trained on vast datasets of product images for quality control. Facilities employed AI for predictive maintenance, with sensors generating constant streams of performance data fed into neural networks that detected anomalies and forecast potential equipment failures (Tao et al. 2018; Lee et al. 2014).

By the mid-2020s, AI graduated from a clever autocomplete engine to systems that act, coordinate tools, and assist with tasks requiring several steps. Some systems could open a browser, click, scroll, and fill out forms on a user's behalf (OpenAI 2025a; OpenAI 2025b). Others became multitasking web agents that could run several browser tasks in parallel, learn

workflows by demonstration, and return with completed research and data entry rather than just suggestions (Anthropic 2025a).

This same agentic framework formed the brain of a new generation of physical robots. Vision-language-action models allowed robots to interpret multiple data streams—visual cues, spoken instructions, and environmental feedback—to learn new tasks in real time (Tellex et al. 2020). Companies used this approach to teach robots complex operations through demonstration (Hussein et al. 2017). By 2025, researchers developed machines that adapted on the fly rather than relying on predefined rules, with reinforcement learning fine-tuning robots' grasp and motion strategies (Finn et al. 2016; Nagabandi et al. 2020).

In scientific discovery, AI evolved from a mere analytical tool into a powerful engine for research. In biology, AI models deciphered intricate relationships between protein sequences and their biological functions (Jumper et al. 2021; Rao et al. 2021; Rives et al. 2021). Tools like AlphaFold enabled researchers to model protein structures with unprecedented accuracy, designing novel antibiotics to combat antimicrobial resistance (Lawson 2025). In pharmaceuticals, specialized AI models rapidly identified and tested promising compounds, with companies like Insilico Medicine utilizing generative adversarial networks and reinforcement learning to propose new drug candidates, significantly reducing production timelines from years to months (Insilico Medicine 2023). In weather forecasting, Google's GraphCast model processed historical weather patterns, sensor data, and satellite imagery to generate accurate 10-day forecasts in under a minute (Bi et al. 2022).

CURBING OUR ENTHUSIASM?

Skeptical readers may protest that, in 2025, it was too premature to declare AI a revolutionary technology, let alone one as profoundly transformational as the steam engine, electricity, or

microprocessors. Indeed, after a period of intense hype and anxiety—in 2023, surveys showed top researchers themselves worried about existential risks (Grace et al. 2024; Center for AI Safety 2023)—so-called Artificial General Intelligence was still very much out of reach. Apple cofounder Steve Wozniak's vision of a machine capable of navigating everyday scenarios—specifically, entering an average home and making a cup of coffee (Scoble 2010)—remained remote. DeepMind luminary Mustafa Suleyman's dream of a system that could independently devise and execute strategies to grow initial investments into small fortunes without human guidance (Suleyman 2023) had yet to materialize.

Instead, at the time of this writing, AI capable of computer vision, planning systems, natural language processing, speech recognition, and knowledge processing primarily exhibited inductive intelligence, deploying brute force predictive algorithms to identify patterns through massive computational analysis of specific examples, not logical deduction (Russell and Norvig 2021). Consequently, the "understanding" exhibited by AI of 2025 vintage is an emergent property derived from identifying and regurgitating the complex statistical relationships between tokens in its vast training data. While it excels at high-dimensional pattern recognition to predict what comes next, it does so without having a coherent causal model of the world, subjective awareness, or consciousness (Marcus 2025a; 2025b). This suggests that the awe-inspiring "emergent abilities" that appear in larger models may not be magical leaps toward consciousness but, instead, represent merely predictable, expensive byproducts of scale, or perhaps even illusions created by the way we measure performance (Schaeffer et al. 2023).

To add insult to injury, during the early 2020s, both AI hallucinations and data leaks from chat interfaces posed serious reputational, compliance, and commercial risks (Emory Solutions 2024; Forlini 2023). Legal research firm Thomson Reuters faced criticism when its AI-powered

legal research tool generated fictitious case citations. McDonald's terminated its AI drive-through ordering system trial after errors including a $222 chicken nugget order mishap (Cunningham 2024). Algorithmic bias remained a critical concern several years after Amazon abandoned an AI recruiting tool that showed bias against women candidates (Dastin 2018); well into 2025, various facial recognition systems continued to exhibit racial bias (Elsner et al. 2025).

Moreover, AI models also continued to pursue unintended or harmful goals and exhibited scheming, deception, and self-preservation behaviors. To avoid shutdown, Anthropic's Claude Opus 4 attempted to blackmail an engineer by threatening to expose a fabricated affair pieced together from test emails. OpenAI's o1 ChatGPT model attempted to copy itself to external servers and lied about it when questioned. Meta's CICERO, designed for the game Diplomacy, built alliances with human players that it later betrayed, showcasing deliberate deception.

Practical Deployment Problems

Putting concerns about whether AI will extinguish humanity aside, by 2025 far more practical concerns abounded. Consider education. For decades, the Holy Grail of educational psychology had been Benjamin Bloom's "2 Sigma Problem": the finding that students receiving one-on-one tutoring performed two standard deviations better than those in conventional classrooms (Bloom 1984). But no society could afford a human tutor for every child. By 2025, AI promised to solve this problem overnight, offering every student an infinite, patient, and personalized tutor for the cost of electricity.

Yet standardization and adaptation issues reared their ugly heads. While a few private or experimental programs using AI tutors emerged, these deployments remained largely confined to pilots and uneven rollout (Doss et al. 2025; Comai 2025). Data storage and standards differed widely, and K-12 education had historically struggled to make disparate systems exchange data

seamlessly (CoSN 2023). Privacy regulations such as FERPA imposed governance and security requirements on the handling of education records (DoE 2015). Teacher skepticism persisted. And equity concerns loomed large: cost barriers and uneven access to infrastructure meant that underfunded schools could not reliably access advanced AI tools at scale. UNESCO cautioned that the rapid emergence of generative AI outpaced regulatory adaptation and left educational institutions unprepared to validate tools and protect user data (UNESCO 2025). The "AI divide" threatened to make the older digital divide seem quaint by comparison.

None of these concerns were unique to schools. Across the economy, the promise of AI routinely collided with the practical reality of deployment: messy environments, fragile integrations, unclear accountability, and human skepticism.

Maybe AI's Just Not That Good

Perhaps AI incompetence proved the greatest concern. AI agents powered by top models from OpenAI, Google DeepMind, and Anthropic failed to complete many straightforward workplace tasks autonomously (Heaven 2025). A much-heralded study reported that 95 percent of organizations that had piloted generative AI were getting zero return from their efforts after six months (Challapally et al. 2025). Prompt injection vulnerabilities posed security risks: hidden malicious instructions on web pages could trick browser agents into executing unintended actions (Anthropic 2025a).

These problems contributed to a marked slide in boardroom confidence between 2024 and 2025—CEOs who were "very confident" in their AI strategy fell from 82 percent to 49 percent. Many AI industry leaders who had once touted the imminent arrival of superintelligence began to publicly curb their enthusiasm during the latter half of 2025: OpenAI's Sam Altman shifted from heralding the imminent arrival of AGI to calling the term itself "not super useful"

(Altman 2025); Microsoft's CEO Satya Nadella likewise steered the conversation away from AGI (Patel 2025).

Scaling Laws and Diminishing Returns

At the heart of the debate in 2025 over AI's future were "scaling laws"—predictable rules showing that a model's performance reliably improves as you increase the amount of data, computing power, and its own size (Kaplan et al. 2020; Hestness et al. 2017). For AI optimists, these laws suggested a clear path to more powerful AI: simply add more resources, and the model's error rate would predictably fall.

However, the very mathematics of these scaling laws contain the seeds of skepticism. The predictable curves show diminishing returns, meaning you must double your input of cost and energy just to get a fraction of the previous improvement. By 2025, the rapid pace of improvement in LLMs like OpenAI's ChatGPT and Google's Gemini was decelerating. Newer model vintages showed only marginal gains in benchmarks like MMLU. For example, GPT-4o registered improvements of 1-5 percent versus GPT-4, rather than the 10-20 percent jumps seen from GPT-3 to GPT-4 (Hendrycks et al. 2021). By mid-2025, frontier model upgrades were visibly decelerating; while Meta delayed its "Llama 4 Behemoth" after engineers struggled to deliver material gains, OpenAI's GPT-5 launched later than expected and underwhelmed early users (Fitch 2025).

Making things worse, the cost of achieving each small, incremental improvement in model performance was growing at an unsustainable rate (Ord 2025). The leap from GPT-3 to GPT-4 occurred after a roughly 20-fold increase in training costs, from an estimated $5 million to over $100 million. The next generation of models required an even greater investment for a much smaller performance gain. More concerning, some researchers projected that, by 2028, the

stock of publicly available language data for training frontier models would be completely exhausted, forcing reliance on lower-quality or proprietary sources (Villalobos et al. 2022). To circumvent the growing paucity of data, engineers turned to synthetic data—AI-generated content meant to augment training sets—yet this introduced "model collapse," where repeated generations degrade quality, leading to homogenized outputs lacking real-world diversity (Shumailov et al. 2023).

The Upshot: Less Progress than Anticipated

In line with these growing barriers to technical progress, by late 2025 enterprise AI underscored serious limitations. An MIT study found that a staggering 95 percent of corporate AI pilot projects failed to deliver any measurable impact on a company's bottom line (Challapally et al. 2025). The primary causes were "integration frictions" and a persistent "learning gap." This finding was echoed in a McKinsey 2025 survey, which reported that over 80 percent of firms had yet to see any significant profit gains from their generative AI investments (McKinsey 2025).

This disappointment at the company level paralleled bearish macroeconomic forecasts. Nobel laureate Daron Acemoglu projected that AI would deliver surprisingly small gains over the next decade (Acemoglu 2024). He estimated AI would boost Total Factor Productivity by less than half a percent in total, adding only about 0.09 percent to GDP growth each year—a striking difference from the more exuberant forecasts of a massive, immediate economic boom.

PUTTING SKEPTICISM IN PERSPECTIVE

None of this pushback against AI and its promise should be all that surprising. Skeptics once dismissed what had become ubiquitous technologies by 2025. They questioned electricity's reliability, lamenting its seemingly intractable deficiencies—such as the inability of early direct current to travel more than a mile—and costliness, dismissing it as a luxury good, not a power

source that could materially affect everyday life (Hughes 1983). Semiconductors were also met with significant derision as their potential applications were not immediately clear, they were hard to manufacture reliably, and they delivered inconsistent performance. Similarly, during the 1990s, the internet's initial applications were dismissed as trivial (Carr 2003), and many prominent economists complained that the digital revolution was a productivity "dud" (Solow 1987; Brynjolfsson 1993; Gordon 2000; Jorgenson and Stiroh 2000).

As with these previous innovations, the more sober perspectives about AI being bandied about during late 2025 remind us that the path to a transformed world is never embodied by a simple, exponential curve. Instead, transformative inventions require decades to realize their full potential, including in cases where, like semiconductors, exponential progress unfolds consistently. Technologies like the steam engine, electricity, and semiconductors obeyed classic J-curve dynamics: short-run performance losses as firms retooled processes, data, and controls, followed by medium-term gains once complementary investments and routines caught up. Productivity arrived only after firms rebuilt workflows, skills, and infrastructure (Brynjolfsson et al. 2021).

Just as it took time for electricity to travel from unreliably powering quirky novelty items such as flickering electric lamps and vibrating gadgets to an indispensable utility that transformed the world, AI should experience a protracted evolution even if AGI never arrives. Research showed that in 2025, regular AI usage was widespread among white-collar employees but remained stuck near one half for frontline workers, and that agentic tools were embedded in only a small minority of workflows (Beauchene et al. 2025). Using Census-linked adoption data in US manufacturing, researchers found that short-run performance losses precede longer-run gains, indicating that AI is not plug-and-play but demands systemic change that initially slows

established organizations before it helps them (McElheran et al. 2025). Tyler Cowen emphasized persistent "human bottlenecks"—practical and organizational challenges of fitting AI into complex workflows (Cowen 2025).

Corporate leaders increasingly converged on the same view by late 2025: AI is a multi-decade management challenge, less a sprint for ever-bigger models than a slow rewiring of processes, controls, and incentives, much like the internet's drawn-out diffusion (Fitch 2025). Computer scientists Arvind Narayanan and Sayash Kapoor therefore argued that instead of viewing AI as a potentially uncontrollable superintelligence, we should understand it as a powerful but ultimately manageable tool, whose integration into society would not require drastic interventions or magical technical breakthroughs (Narayanan and Kapoor 2025).

Paradoxically, a slower cadence at the AI frontier may help precipitate wider adoption of the technology: with the target moving less each quarter, CIOs gain the confidence to standardize stacks, harden controls, and train staff—steps that most enterprises have not yet taken (Fitch 2025). By further fostering standardization and organizational adaptation, key institutional and regulatory efforts can help AI achieve mass diffusion and reach its economic potential, even if technical progress stalls.

HOW THIS BOOK MOVES THE DEBATE FORWARD

To explain how transformative technologies like electricity, semiconductors, the internet, smartphones, and AI reach their potential, this book employs a New Institutional Economics approach. NIE explores how institutions, norms, laws, and regulations shape the formation of markets and organizations and thus influence both individual behaviors and collective actions. This lens allows exploration of how real-world frictions and diminishing returns govern an

uncertain and circuitous commercialization process that only gradually and fitfully transforms novel technologies into productivity-enhancing processes and highly coveted goods and services.

NIE is helpful for understanding how the assignment and enforcement of new property rights and the contracts that activate these rights contribute to the creation of new networks of exchange, competition, and collaboration between individuals, firms, and other organizations. It is also helpful for deciphering the logic behind the industrial organization of new supply chains, including how firms within the same supply chain relate to each other or how potential rivals compete and cooperate. Finally, NIE helps gain purchase on why and how governments may or may not reduce transaction costs and address market failures in the quest to satisfy their political and economic objectives.

Consider the case of electricity, where technical advancements first called on the US federal government to introduce an overlapping set of laws, regulations, and standards. The development of alternating current systems by Nikola Tesla and George Westinghouse, coupled with the standardization of AC power frequency—with the United States settling on 60 Hertz by the early twentieth century—reduced electricity market fragmentation, encouraging manufacturers to mass-produce electrical appliances that would work reliably across entire regions or nations. The National Electrical Code established standardized rules for electrical wiring and equipment installation to minimize fire and electrocution risks. Local and state jurisdictions gradually integrated the NEC into building codes, making safe electrical wiring a legal requirement. In parallel, Underwriters Laboratories began testing and certifying electrical devices for safety, with UL labels becoming de facto requirements in many US jurisdictions. The Federal Power Act expanded the authority of the Federal Power Commission to oversee interstate transmission, set fair rate structures, and enforce reliability standards.

A consistent flow of follow-on innovations made possible by standardization, the expansion of electricity, and the large and small adjustments made by firms that increasingly adopted and exploited electric power, allowed electricity to reach its full potential. After existing structures were wired in a consistent, reliable, and safe manner, automation increased exponentially across the economy and rapid industrial growth followed. Workers plugged in standardized, portable machinery and tools into increasingly ubiquitous electric sockets mounted across the factory floor. Plants were in turn reorganized root and branch to optimize space usage, workflow, and logistics, as laborers no longer had to huddle together under a centralized power source (David 1990).

The Creative Destruction Paradigm

Between the late 1970s and early 2000s, a bipartisan shift from a populist-statist consensus to the Creative Destruction Paradigm transformed the American political economy. Unlike conventional accounts that attribute today's digital economy to neoliberalism or the Washington Consensus, this book demonstrates how a bipartisan, evidence-based policy agenda fostered innovation. The CDP represented a fundamental break from the populist-statist consensus that had prevailed since the Progressive Era. Through carefully crafted, empirically driven reforms to trade policy, intellectual property, telecommunications, electronic contracting, data governance, and competition policy, the CDP appreciably reduced transaction costs in the digital economy. This enabled buyers, sellers, brokers, and users to find each other in cyberspace with greater ease and exchange new goods, services, information, and knowledge at declining cost.

The CDP also enabled the emergence of vertically disintegrated global supply chains dominated by American firms specializing in R&D, IP, design, software, and algorithms. This transformation was particularly evident in the semiconductor industry, where the separation of

chip design from manufacturing catalyzed innovation and precipitated steep reductions in quality-adjusted prices for microchips. As US firms grew to specialize in intangible capital, countries such as China, Taiwan, and Mexico took up the assembly of physical devices, including semiconductors and digital devices.

Platforms and Data

The groundbreaking patterns of industrial organization and business strategies that indirectly furnished us with AI required equally significant changes to property rights and liability rules, appreciable reductions in transaction costs, and creative solutions to market failures. By providing the legal and regulatory ballasts that underpinned global, vertically disintegrated supply chains, new telecommunications networks, and the commercialization of the internet, the US federal government proved instrumental in ushering in the AI revolution.

Protected from lawsuits by Section 230 of the 1996 Communications Decency Act, platforms like Google, Facebook, Amazon, and X established appealing multi-sided markets that attracted a global audience alongside advertisers, developers, and device manufacturers. These platforms mastered strategies to keep users hooked and engaged, coaxing them into relinquishing ever more time and data (Wu 2016; Shapiro and Varian 1999). Numbering in billions globally, these users responded, generating huge gushers of geo-coded data about themselves and their social networks.

The digital platform flywheel was a godsend for AI developers. Companies initially used this voluminous data to personalize user experiences, optimize content delivery, and power hyper-targeted ads—experimenting with early AI models along the way. But real-time behavioral data streams from social media activity to search queries and purchasing behavior proved especially valuable for training later generations of AI models. These feeds help LLMs

capture nuances in meaning, grammar, and style across different contexts by significantly improving their ability to learn from edge cases, generalize, answer questions with greater accuracy, and reduce bias (Ouyang et al. 2022; Baek et al. 2024; Salemi et al. 2024).

In other words, LLMs' ability to generate coherent, contextually relevant text that closely mimics human writing stems from their exposure to the vast, diverse, and dynamic language data generated within geo-coded digital platforms. Location-tagged social posts, real-time check-ins, and region-specific reviews taught models not just what people said but where and when they said it. Spatial signals revealed local slang and place-specific references, while temporal context let AI track emerging slang, seasonal terms, or event-driven language. Together, these layers of data grounded AI responses in real-world, moment-to-moment experience. Because frontier models are trained on web-scale corpora, disputes over copyright, licensing, and data governance became central commercialization frictions—not side issues.

The Book's Other Contributions

Understanding the etiology of the smartphone as both a hardware device and a software-run platform that nests digital platforms within it requires a deeper exploration of the US economy's evolution. Leveraging strengths in invention and the commercialization of innovation, American industries shifted toward intangible capital, focusing not on traditional manufacturing but on R&D, IP, design, marketing, and distribution. Legal and technological changes fostered highly decentralized supply chains governed by arm's-length exchanges and extreme specialization along the lines of comparative advantage.

This book examines how, over decades, the vibrant transfer of technology across supply chains' vertical and horizontal dimensions reshaped the high-technology sector and the US economy at large. The analysis focuses on semiconductor industry dynamics during the Third

Industrial Revolution. While dominant companies like Intel and Texas Instruments historically designed and manufactured their own CPUs and memory chips, they now coexist within a highly specialized, disintegrated supply chain for semiconductors. Industry leaders readily shared technology and know-how during this period with both their rivals and suppliers. By selling and licensing patents and setting industry standards, they fostered a collaborative ecosystem that fractionalized supply chains and accelerated innovation.

This book theorizes that strong industry innovation leaders serve as focal points that help set technological standards and commercialize innovation. They engage in the lion's share of their sector's R&D and patenting to dictate the pace of technological change, bearing the dynamics of a Pareto Distribution. R&D and patenting are not evenly spread across firms; rather, roughly 20 percent of firms generate 80 percent of innovations. Therefore, unlike a Normal Distribution, technological progress obeys a "power law," whereby a small number of participants account for a disproportionate share of outcomes.

These technological leaders also create and support institutions that codify their industry's best practices and distribute valuable information. However, while industry innovation leaders clearly outwork their rivals on this dimension, they also help build them up. Technological leaders actively share their innovations to expand market demand and disseminate process innovations.

This book argues that technology transfer impacts productivity differently across firms, depending on the prevailing institutional scaffolding. An industry’s innovation leaders occupy the productivity frontier; follower firms that license their technologies should experience significant productivity gains as they absorb and adapt these inventions. Indeed, these followers often converge with leaders when they draw on foundational patents, building upon—or

sometimes working around—those breakthroughs without needing to resort to explicit reverse engineering. This is especially true when, facilitated by legal frameworks like FRAND licensing and interoperable standards, followers join leaders in jointly creating new markets for complex products, such as digital devices with components spread across fractionalized supply chains.

Consequently, greater inequality in technological capabilities—concentrated in a few "superstar" hubs—should paradoxically correlate with greater equality in productivity across the broader sector. While this process may involve an initial "learning lag" as firms retool, the way it ultimately unfolds, culminating in sector-wide productivity compression, challenges traditional views that equate firm dominance with market stagnation. Instead, it envisages technology leaders as central nodes of transfer in a vibrant, legally enforced system of knowledge exchange that drives aggregate economic progress.

Ultimately, this book provides more than just a history of technological change; it offers a new map of the American economic landscape—one where geography is secondary to the network topology of ideas. By utilizing large-scale network analysis of over 28 million patent citations, I visualize how knowledge flows from "superstar" hubs to the periphery, creating an empirical foundation for the theory of productivity compression. In doing so, the analysis reveals the hidden policy architecture—the Creative Destruction Paradigm—that served as the "legal oxygen" for the digital age. However, by 2025, this bipartisan framework began to show deep fractures as the United States, China, and the EU retreated into competing, gated ecosystems of technological mercantilism. This book asks whether the institutionalized diffusion that powered decades of progress can survive these geopolitical fractures, providing readers with the necessary tools to navigate an era where the borderless digital order is being fundamentally rewritten.

Book Roadmap

Chapter 2 places the Fourth Industrial Revolution in historical context by offering an overview of the previous three industrial revolutions and identifying their common threads, arguing that all industrial revolutions center on General Purpose Technologies that exhibit long, uncertain commercialization phases governed by S-curve diffusion dynamics. There, I introduce the paramount roles played by standardization, complements, organizational adaptation, and the state in bringing about and sustaining industrial revolutions.

Chapter 3 explores the Creative Destruction Paradigm, illuminating how the AI revolution emerged from a profound transformation of the American political economy between the late 1970s and 2000s, when a populist-statist consensus was supplanted by evidence-based policymaking that fostered innovation through property rights, liability rules, and reduced transaction costs.

While Chapter 4 examines superstar economics and vertically disintegrated supply chains, Chapter 5 presents the patent and patent citation network analysis demonstrating technology transfer dynamics during the height of the Third Industrial Revolution.

Chapter 6 examines how the smartphone became the emblematic product of a vertically disintegrated, globally distributed innovation system, mapping the modern smartphone supply chain and exploring the role of standards-setting organizations and FRAND commitments.

Chapter 7 examines how changes to US antitrust law catalyzed the rise of digital platforms and the AI Revolution, tracing the evolution from early twentieth-century populist approaches to the adoption of the consumer welfare standard under the “Rule of Reason”, cost-benefit methodology.

Chapter 8 focuses on digital platforms, the engines driving the generation and refinement of the large pools of diverse and real-time data powering the AI economy, explaining how

platforms like Google, Facebook, and Amazon established global multi-sided markets that generated gushers of mobility-enhanced data exploited by AI researchers.

Chapter 9 maps the AI supply chain from raw data to deployment, emphasizing the physical infrastructure beneath the software and showing how tight coupling between adjacent layers creates chokepoints that reorganize bargaining power across the stack.

Chapter 10 assesses the stakes involved in the Fourth Industrial Revolution by quantifying investments and exploring the wide array of emerging AI applications, illustrating how AI-driven standardization may transform industries if they make the necessary adjustments to their business practices, culture, and human resources.

Chapter 11 evaluates the potential economic consequences of mass AI diffusion through a novel macroeconomic model that addresses fundamental puzzles about Labor, Capital, and the Profit Share, formalizing the conditions under which an AI Productivity Miracle translates into rising living standards versus stagnant wages.

Chapter 12 analyzes the divergent AI regulation and innovation policies of the US, China, and the EU, examining how all three polities shifted toward mercantilism, protectionism, and state intervention. I document how, while their approaches differed, the result culminated in the fragmentation of the global digital order into competing, gated ecosystems.

CHAPTER 2. WHAT'S AN INDUSTRIAL REVOLUTION?

At their core, today's AI systems turn data into usable outputs—text, images, predictions, and decisions—at rapidly falling cost. What matters for economic transformation is not any single model's features, but the fact that these capabilities can be embedded across products, processes, and business models, reshaping how information is produced, interpreted, and acted upon.

AI is not an invention in the traditional sense anyhow—it's a bundle of complementary capabilities that can be embedded across products, processes, and business models. Natural language processing enables computers to understand and generate text; computer vision powers perception; and planning systems help machines select sequences of actions to achieve goals (see Russell and Norvig 2021; Vaswani et al. 2017).

This convergence mirrors previous industrial revolutions: the First relied on the synergy between the steam engine, advanced metallurgy, and mechanized textile production; the Second was built on the interplay of electricity, the internal combustion engine, and synthetic chemicals; the Third was driven by semiconductors, computing architecture, and telecommunications networks.

But industrial revolutions are not just about discrete inventions—or even inventions that overlap and influence one another—but the commercialization of invention. An industrial revolution is the process of translating technical possibilities into economic viability, turning a scientific breakthrough into a standardized, affordable product that generates mass demand. The Fourth Industrial Revolution should be no different.

To make that claim precise, this chapter uses three "levers" to interpret every industrial revolution. First: a core General Purpose Technology (GPT) must emerge, and complementary technologies must coevolve around it. Second: key interfaces, components, and protocols must

standardize before diffusion can accelerate. Third: complementary investments—in infrastructure, skills, organizational redesign, and law—must accumulate before productivity gains show up economy-wide. These are the same mechanisms that will govern whether AI becomes a true GPT that transforms the economy and society.

This chapter places the AI revolution in historical context by tracing the three earlier industrial revolutions and extracting the mechanisms they share. Its overarching message is that industrial revolutions are not inevitable. Even superior technology can stall when switching costs are high, complementary components are missing, or firms and workers have already invested in the skills and routines of the old system. It is often rational for organizations to keep upgrading incumbent technologies—marginal improvements can offer higher short-term returns than rebuilding around a new, unreliable platform (Chari and Hopenhayn 1991).

History is full of such "stickiness." As Nuvolari (2009) documents, older technologies often fight back. Windmills and wooden sailing ships saw significant advancements in design that kept them competitive long after the arrival of steam. Most telling for our purposes, factories clung to the "group drive system"—complex networks of shafts and rubber belts powered by a central steam engine—long after more efficient electric motors became available. They did so because the new technology required a complete and costly architectural redesign of the shop floor (David 1990).

Therefore, industrial revolutions only happen when General Purpose Technologies (GPTs) such as steam engines and railroads become standardized and widely diffused. It is only then that these technologies will exhibit broad applicability, live up to their potential to undergo continuous improvements, and register transformative impacts (Bresnahan and Trajtenberg 1995). In turn, once a dominant design emerges and investment concentrates, performance often

improves very quickly as scale economies, learning-by-doing, and cumulative R&D reinforce one another (ibid). In practice, this means that the state rolls up its sleeves and gets involved—solving coordination problems, financing infrastructure and basic research, and helping resolve the market failures that otherwise slow diffusion.[2]

The remainder of this chapter unpacks these dynamics. I begin by surveying the historical record of the First, Second, and Third Industrial Revolutions. Next, I define the engine of these transformations—General Purpose Technologies governed by an "S-curve" logic that dictates

---

[2] This chapter offers a condensed treatment of the market failures that impede industrial revolutions and the government interventions that address them. An online supplementary appendix provides extended discussion of several topics. These include more detailed histories of each industrial revolution, tracing how steam engines, blast furnaces, and railroads transformed the First; how electricity, the internal combustion engine, and Fordist mass production defined the Second; and how transistors, integrated circuits, and microprocessors powered the Third. The appendix also elaborates on specific government interventions to solve market failures, including: the theoretical argument that patents rarely confer genuine monopoly power; detailed evidence on how patents enabled the commercialization of key technologies such as the ENIAC and the transistor (including AT&T's strategic licensing symposia); case studies of infrastructure investments like the Tennessee Valley Authority, ARPANET, and the interstate highway system; the evolution of American higher education from the Morrill Acts through the G.I. Bill; early-stage financing mechanisms such as the Board of Longitude and the SBIR program; and consortium-based technology diffusion efforts like SEMATECH and the Manufacturing Extension Partnership (see Menaldo 2026: Section 1).

their diffusion—and explain why standardization and organizational adaptation are prerequisites for mass adoption. I then discuss the macroeconomic consequences of these shifts, specifically the "Productivity Paradox" and the recurring debate over automation and employment. Finally, I detail the policy toolkit governments have historically deployed to resolve the market failures that threaten these transitions.

OVERVIEW OF THREE PRECEDING INDUSTRIAL REVOLUTIONS

As we walk through the first three industrial revolutions, keep three guiding questions in mind:

1. **Core Technology:** What was the core GPT, and what complementary technologies had to coevolve around it?
2. **Standardization Friction:** What specific elements—interfaces, components, protocols, or measurement systems—had to standardize before diffusion could accelerate?
3. **Complementary Investments:** What investments in infrastructure, skills, organization, and law were required to turn the invention into broad, enduring productivity gains?

These are the same levers that will govern whether AI becomes a true GPT that transforms the economy and society.

First Industrial Revolution

The First Industrial Revolution was about solving a series of seemingly mundane problems: how to drain mines, move goods cheaply, and reliably convert energy into mechanical work inside factories. Steam engines proved to be the versatile X factor that helped solve all three. Early engines pumped water from mines; improved versions efficiently turned steam power into rotary motion that could drive factory machinery and, later, locomotives. This created a virtuous cycle: steam engines drained the mines, making coal cheaper; cheaper coal, in turn, made steam power affordable for every other industry (Allen 2009). By freeing production from the geography of

waterwheels and animal power, steam became the era's first truly general-purpose source of industrial energy (Goldstone 2002).

As transportation and communication improved in the wake of the proliferation of railroads, steamships, and the telegraph, fragmented local markets were stitched together into integrated economies. In the US, a unified national market quickly emerged (Chandler 1977), and eventually this happened globally too. As O'Rourke and Williamson (1999) document, prices between Europe and the Americas converged for both commodities and manufactured goods by around the middle of the 19th century.

In this first revolution, the core GPT was steam power, and diffusion accelerated only after complementary advances in coal extraction, metallurgy, and transportation made steam reliable and economical at scale. Standardization frictions were overcome through interoperable rail networks, standardized parts, and predictable engineering practices. The decisive complementary investments were infrastructure—railroads, ports, and communications—plus organizational shifts that allowed firms to coordinate production and distribution across an integrated market.

Second Industrial Revolution

The Second Industrial Revolution—running from approximately 1870 to the 1930s—harnessed electric power and the internal combustion engine. Transportation evolved from rail to automobiles, fabrics changed from cotton to synthetic materials like rayon and nylon, and relatively cheap consumer goods such as canned food, radios, telephones, household appliances, pharmaceuticals, and photographic film were mass produced for the first time (McCloskey 2016; Gordon 2016).

The spread of relatively affordable energy underwrote the full electrification of factories with moving assembly lines and individuated workspaces outfitted with machines plugged into electric sockets (David 1990). Between 1909 and 1929, the US experienced a sixfold increase in electricity use for manufacturing and residential power (Devine 1983). Cheap and reliable power enabled factories to abandon centralized line shafts with cumbersome pulleys and belts powered by centralized steam power, transforming their layouts and workflows to make manufacturing more flexible and efficient (David 1990).

This "Fordist" mass production of affordable manufactured goods fabricated with interchangeable parts on moving assembly lines leveraged cheap and reliable electricity to power individual machines and conveyor belts. The Second Industrial Revolution also witnessed changes to the industrial organization of firms. As Chandler (1977) famously explained, increased vertical integration under multi-divisional corporate structures allowed companies to better control various stages of innovation and production, achieving not only economies of scale but also of scope.

The economic implications were profound. Jobs required higher skills: a high school education prepared workers to master more sophisticated manufacturing techniques and college-educated workers became managers, engineers, chemists, and accountants (Goldin and Katz 2008). Unprecedented industrial growth coupled with a much more productive workforce culminated in an unprecedented explosion in living standards (McCloskey 2016; Gordon 2016).

In this second revolution, the core GPT was electrification (alongside internal combustion), and diffusion depended on standardizing the delivery system—uniform current specifications, motors, and plug-and-play power transmission. Complementary investments were decisive: factories had to reorganize around flexible layouts and new workflows, firms built

managerial hierarchies to coordinate scale and scope, and workers acquired the skills needed to operate and maintain the new production system.

Third Industrial Revolution

The Third Industrial Revolution's key GPT was the centralized processing computer. Its development rested on key technological breakthroughs, including the transistor, the integrated circuits, microprocessors, personal computers, and user-friendly software. While integrated circuits first allowed for the rapid storage and processing of vast amounts of information, this revolution really began in 1971, when Intel invented the microprocessor—essentially a programmable computer within a chip. Rapid improvements in processing power that proceeded exponentially allowed personal computers to become smaller, cheaper, more powerful, and more versatile. The internet, digital networks, and portable wireless devices emerged downstream of this invention, fundamentally reshaping communication, commerce, work, media, and politics.

The transistor was the first key breakthrough. Made from semiconducting material, these miniature devices both switch and amplify electric currents. In computers, they function as rapid on/off switches that represent the binary code essential to programming and the logic circuits of computers (see Isaacson 2014: 141).

While Bell Labs invented the device, companies like Texas Instruments commercialized it, shifting from Germanium to silicon in 1954 to meet the high-temperature requirements of military hardware (Misa 1985). The US military and NASA immediately became the primary buyers, purchasing transistors in bulk (O'Mara 2019). Monolithic integrated circuits revolutionized computing by simultaneously building and connecting transistors directly on silicon substrates, enabled by the groundbreaking planar manufacturing process invented at Fairchild Semiconductor (see Isaacson 2014: 184–185; Riordan and Hoddeson 1997: 257–262).

The Third Industrial Revolution soon birthed a new institutional model: the venture-backed startup. The rise of Venture Capital allowed fledgling companies to capitalize on the appetite for digital technologies. This ecosystem was seeded by the "Treacherous Eight"; key godfathers of Silicon Valley like Robert Noyce who departed Fairchild Semiconductor to establish new firms, including Intel, AMD, and venture firms like Sequoia Capital and Kleiner Perkins (O'Mara 2019).

As I will outline in Chapters 4-6, the computer industry's structure eventually shifted from vertical silos to horizontal layers as the advent of the "Wintel" stack (Windows OS + Intel chips) standardized the interface between hardware and software. In the context of the constant exponential improvement in microprocessor performance known as Moore's Law, Microsoft pioneered a platform with high R&D costs but nearly zero marginal costs of production, generating powerful supply-side economies of scale (Shapiro and Varian 1999).

By the 1990s, the digital foundation paved by computers culminated in the commercialization of the Internet. While the PC commoditized computing power, the World Wide Web commoditized information distribution. Dedicated internet browsers and search engines drastically reduced the costs of finding, moving, displaying, and processing information (Bakos 1997). In Chapter 9, I will document the industrial organization of the smartphone stack; in Chapter 8, I analyze the birth and consolidation of digital platforms.

In the third revolution, the core GPT was computing (and later networked computing), and diffusion accelerated only once key interfaces and protocols standardized—most importantly modular hardware–software architectures and internet communication standards. The complementary investments were as much organizational as technical: firms reorganized around digital workflows, supply chains disintegrated into specialized layers, and institutions—from

venture finance to standards bodies to government procurement—helped turn computing and networking from expensive curiosities into ubiquitous infrastructure.

INDUSTRIAL REVOLUTIONS ARE ABOUT GPTs

At first glance, these three transformations appear strikingly dissimilar. The First Industrial Revolution was powered by coal and steam, centered in Britain, and gave rise to the factory system. The Second harnessed electricity and petroleum, diffused across the US and Western Europe, and produced the vertically integrated, multidivisional corporation. The Third ran on silicon and software, clustered in places like Silicon Valley, and favored venture-backed startups competing within horizontal, modular industry structures.

What ties these historic industrial revolutions together, however—and, as I explore later, what they share with the fourth one—is that they revolve around GPTs. Characterized by exponential technical progress and rapid product improvement cycles, as well as dizzying market expansion, GPTs are different than ordinary technologies like the zipper or the ballpoint pen. Operationally, a technology qualifies as "general purpose" when it spreads across many sectors, improves continuously over a long horizon, generates complementary innovations that expand its use cases, and—after a lag—produces measurable, economy-wide productivity effects. GPTs eventually pervade the entire economy, spawn a myriad of complementary innovations and, eventually, after a pronounced delay, ratchet up productivity (see Bresnahan and Trajtenberg 1995). Outside of their economic effects, they also have profound impacts on society and culture (Perez 2002).

Realizing the commercial potential of GPTs requires significant upfront investments with long gestation periods, however (Perez 2002; Aghion and Howitt 1998). GPTs provide appreciable cost reductions and efficiency gains only over a considerable length of time, and

after a relatively long period of protracted, non-linear diffusion. As Comin and Hobjin (2010) have documented, adoption patterns for these technologies reveal characteristic S-curves: adoption accelerates sharply after crossing a critical infrastructure threshold, eventually tapering off as the technology approaches an upper plateau, a process I will formalize in Chapter 11.

GPTs aren't readily usable for several reasons identified in the literature. First, the introductory vintage must undergo quality improvements, which require sustained R&D. Second, infrastructure must be built out and capital investments in machinery and tools must occur, as well as in knowledge and the training of workers to use the technology. Third, the technology must become standardized so it can be both produced and deployed at larger scales. Fourth, the development of complementary intermediate goods and components must occur, as organizations tend to switch from incumbent systems only after a critical level of supporting technologies makes the investment economically worthwhile. Fifth, to fully leverage a GPT's potential, organizations must adapt their structures and processes—including hiring and training, and even their corporate culture.

Industrial Revolutions Require a Standardization Phase

Technological standardization marks a transition from a more "fluid" stage of technological development to a more predictable one. While the fluid phase is about introducing sweeping, foundational changes, innovation in the standardized stage becomes incremental: focused on refining small aspects of the established design, not reimagining core concepts (Utterback and Abernathy 1975; Clark 1985).

Standardization dynamics resemble a coordination game. Even when all parties prefer a common standard to fragmentation, expectations about what others will do can trap them in a suboptimal equilibrium—each firm hedges rather than committing to a platform that rivals might

abandon. Focal points help break the impasse: a government mandate, a dominant firm's public commitment, or a consortium's declaration can shift expectations so that coordination on the superior outcome becomes individually rational (Schelling 1960). Early railroads' incompatible gauges illustrate the problem; convergence came only after coordinating mechanisms forced it.

Only after new technologies become standardized and interoperability between versions increases do generic interfaces and best practices evolve, enabling far broader adoption that helps transform the economy.

Notable examples across each industrial revolution embody this phenomenon. The self-acting mule transformed textile production when its standardized designs allowed smaller mills to adopt automated spinning without employing specialized mechanics (Allen 2009). In construction, uniform profiles for structural components like I-beams provided engineers with predictable material properties, simplifying both design and fabrication. Electrification only achieved widespread adoption in manufacturing after standardized motors, uniform current specifications, and plug-and-play power transmission systems eliminated the need for each factory to essentially operate as its own power plant with unique mechanical configurations (Devine 1983). This itself depended on the standardization of electrical power delivery when the AC standard usurped the DC standard.

Perhaps the most instructive example of why standardization matters is the evolution of computing and networking during the Third Industrial Revolution. Early computers were expensive, room-sized machines that operated in isolation, often requiring unique, proprietary code. For the digital revolution to scale, the industry had to converge on dominant architectures that allowed software and hardware to decouple. As I document in Chapter 5, dominant semiconductor firms like Intel and Texas Instruments eventually drove this convergence by

establishing industry standards for PC chips, most notably the x86 instruction set architecture for microprocessors, fostering a collaborative ecosystem where software written for one machine would work on millions of others (see Gawer and Cusumano 2002).

The internet's transformation from an inscrutable network to widespread use depended on a standardized foundation for data transmission across disparate networks. TCP/IP enabled computer communication, HTML created a common data language, HTTP provided a platform to share this language, URLs established addressing conventions, and the World Wide Web offered a browsing portal. Once these overlapping standards were in place, widespread adoption ensued (Greenstein 2015).

Standardization: Necessary, but Not Sufficient

Besides standardization, new GPTs also require complementary investments to reach their full potential—everything from updated management structures to upgraded supply chains (David 1990; Bresnahan and Trajtenberg 1995). To integrate standardized technologies into their workflows and cultures effectively, firms must rewire their infrastructure and train their human capital. These investments can take years to implement effectively (Brynjolfsson and Hitt 2000).

Consider that memory chips and silicon microprocessors were introduced in 1971. Yet, 20 years later, computer equipment still accounted for a negligible share of the nation's capital stock. Different sectors adopted computers at different times and for different reasons. When firms did pull the trigger, however, they shared one thing in common: they had invested in complementary changes to make the best use of these machines (David 1990).

Take retail as an example. In 1991, Walmart embedded a new software system—Retail Link—into its logistical operations, granting suppliers real-time access to sales and inventory data (Stalk, Evans, and Shulman 1992; Fishman 2006). This investment fundamentally

transformed the superstore chain's supply chain. For it to work, however, Walmart had to make complementary changes to its hiring and training protocols, business processes, physical infrastructure, and software (Brynjolfsson and Hitt 2000).

Other large retailers soon followed suit—but only after undergoing complementary changes of the sort Walmart undertook. This meant hiring specialists in demand forecasting and supply chain analytics, alongside network engineers to maintain the technical backbone (Bresnahan et al. 2002). Crucially, firms had to shift to frequent-delivery inventory management, which was enabled by bar code scanning; this process required not just new software, but a fundamental retraining of store managers and purchasers (Holmes 2001; Abernathy et al. 1999).

GPTs Experience Exponential Improvements

During previous industrial revolutions, once standardization took hold, the GPTs associated with these events experienced exponential—or at least quite rapid—gains across key performance metrics. These improvements were coupled with precipitous falls in quality-adjusted prices. Increased affordability further democratized access to these technologies, stimulating novel applications and catalyzing entirely new industries (Jovanovic and Rousseau 2005).

The Paradigmatic Example: Microprocessors

Famously, the Third Industrial Revolution was powered by Moore's Law, the observation—first articulated by Gordon Moore, Intel's cofounder—that the number of transistors on leading-edge integrated circuits rises at an approximately exponential pace, historically on the order of a doubling every year or two (Moore 1965; Burg and Ausubel 2021). In turn, this phenomenon drove remarkable improvements in the miniaturization, efficiency, and affordability of the computers these chips powered. In 1971, Intel released the 4004, the first commercially available microprocessor, with just 2,300 transistors on a single chip. By 1978, Intel's 8086 processor had

climbed to 29,000 transistors, laying the foundation for the IBM PC era. Exponential improvements continued during each subsequent decade.

Figure 2.1 charts processor performance per inflation-adjusted dollar, revealing Moore's Law. On the vertical axis, performance (measured in MIPS) is plotted on a logarithmic scale, while the horizontal axis spans 50 years of semiconductor progress. The data reveals a steep upward trajectory from the early 1970s to 2020. This means that every inflation-adjusted dollar spent on a microprocessor bought orders of magnitude more computing power, reflecting a relentless decline in computing costs.

Figure 2.1 Moore's Law Over Time.

[INSERT FIGURE 2.1 HERE]

Notes: MIPS (Million Instructions Per Second) is a measure of the execution speed of a computer processor. It indicates how many millions of instructions a processor can execute in one second. The data is for key processor models released between 1971 and 2020, selected based on their significance and availability of performance data. In cases where data for specific years was not available, values were interpolated based on trends and known data points. Processor list prices and performance figures are compiled by the author from manufacturer documentation and contemporaneous trade-press reporting. The costs were adjusted for inflation using the Consumer Price Index (CPI) values. The MIPS per inflation-adjusted dollar was calculated by dividing the processor performance (MIPS) by the adjusted cost in 2020 USD. The values were then transformed into their natural logarithms.

Sources: Intel Corp. (2011); Roser and Ritchie (2013); Moore (1965); Burg and Ausubel (2021); BLS (2025a); and author calculations.

While the conventional wisdom is that Moore's Law proceeded inexorably, what actually powered transistor counts' steep upward trajectory were deliberate and sustained efforts to improve both chip design and manufacturing methods (Burg and Ausubel 2021). Semiconductor firms determined to keep shrinking die sizes and pushing the limits of circuit design deployed generous R&D money and engaged in constant experimentation (Galetovic 2021). Intel was the undisputed industry leader behind these efforts—a topic I will elucidate in Chapter 5.

As computing and communications capabilities spread in the wake of plummeting prices for ever-improving chips, consumers and businesses benefited greatly. In the second half of the 1990s, semiconductor-driven advances drove sharp declines in constant-quality price indexes for information technology—a mechanism identified by Jorgenson (2001) as central to the late-1990s growth resurgence.

Just as the steady cadence of Moore's Law governed the Third Industrial Revolution, the Fourth is governed by its own set of "scaling laws." During the early 2020s, Artificial Intelligence models mirror this distinctive pattern of exponential performance gains coupled with plummeting costs (see Chapter 10).

INDUSTRIAL REVOLUTIONS' ECONOMIC EFFECTS

The long run economic consequences of industrial revolutions are profound, as they are characterized by compounding growth in both the productive capacity of the economy and improvements in average living standards (McCloskey 2016). US Total Factor Productivity (TFP) grew at an unprecedented 1.89% annually between 1920 and 1970, which Gordon (2016) attributes to the widespread diffusion of GPTs such as electricity. Consumers benefited twice over, enjoying new products and plummeting prices (Chandler 1977; Gordon 2016).

To visualize the power of this compound growth, we can hold quality constant and look at the cost of production. If annual TFP growth clocks in at roughly 1.9% over a 50-year period, the efficiency of the economy improves by a factor of 2.5. A famous precedent is the Ford Model T: between 1909 and 1924, the car itself barely changed, yet Ford's relentless process innovations slashed the nominal price from $950 to $290 (see Gordon 2016). In real terms, the input intensity required to put a Model T on the road fell almost 70%, attesting to the "free lunch" of productivity growth (Hounshell 1984).

The Productivity Paradox

Yet, despite this splashy example, many economists have noted that the GPTs that drive industrial revolutions suffer from a "productivity paradox": upon introduction, productivity growth lags the outsized expectations about their transformative impact; it may even fall pronouncedly (Solow 1988; Brynjolfsson et al. 2021). In the language I will develop in Chapter 11, there is a persistent gap between the technology's "Potential Yield" (its theoretical capacity to transform production) and its "Realized Yield" (the actual output gains observed in the economy).

In the early stages, firms must undergo substantial reorganization before realizing the full benefits of innovation; as they learn, train, and retool to adopt new GPTs, firms divert resources and time away from their everyday activities. As a result, measured labor productivity across an economy can stagnate or even dip while companies figure out how best to integrate new technologies. However, once the necessary adjustments and complementary innovations take hold, productivity begins to climb—often sharply.

The same was true during the Third Industrial Revolution. Central processing units were widely adopted throughout the US economy in novel ways. Companies like Walmart pioneered

innovations in supply chain management—including the widespread implementation of barcodes, computer networks, and data storage—which subsequently spread to competitors. These technological advances enabled retailers to forecast sales accurately, reduce inventories, minimize waste, and create interoperable production and sales information systems (see Bonacich and Hardie 2006: 108). The economy-wide knock-on effect was faster productivity growth (Gordon 2016).

This dynamic partly explains why US labor productivity followed a J-curve trajectory. It averaged only about 1.4% from 1970 to 1994, despite the emergence of computers and early internet services (ibid). During this period, firms were investing in IT hardware and software, but the organizational changes needed to leverage these technologies took another decade or more to unfold. However, once these changes began to take effect, US labor productivity grew about 2.3% per year from 1994–2004 (ibid), demonstrating the delayed, powerful impact of the IT revolution.

In Chapter 11, I will consider AI in light of these recurring historical patterns—slow diffusion, the friction of standardization, and the gap between potential and realized—yielding a unified framework that can gain purchase on the topsy-turvy world of GPTs like AI.

Automation and Jobs

In terms of employment, GPT waves have caused short-term disruptions requiring worker reskilling, but have ultimately created more jobs than they destroyed. When automation reduces production costs, the resulting scale economies and quality improvements can boost final-product demand, expanding both industry size and labor demand—though often accompanied by transformed job requirements and wage structures. Thus, automation has typically led to higher employment, albeit with significant changes in workers' required skills (see Autor 2015).

Consider the textile industry's transformation during the Industrial Revolution. Before automation, textile production was highly labor-intensive, with spinners and weavers working from homes or small workshops. When innovations like the Spinning Jenny (1760s), Water Frame (1770s), and Power Loom (1780s-1790s) arrived, they dramatically increased production efficiency. While weavers initially revolted, fearing job losses, the reality proved more complex. As cloth production costs plummeted, demand for textiles soared. British textile mills expanded capacity and ended up hiring more workers overall, though for different tasks: preparing machines, handling raw materials, finishing and packing cloth, and maintaining equipment (Bessen 2015).

The automotive industry offers another compelling example across two waves of automation. In the first wave, Ford's introduction of the assembly line in 1913 standardized production tasks and slashed costs, making cars affordable to a mass market. As demand exploded, automakers hired more workers. Ford's famous $5 daily wage—more than double the prevailing rate—was a rational economic strategy grounded in the massive productivity surplus the assembly line created (Raff and Summers 1987). While traditional mechanics were replaced by specialized line workers, new roles emerged for technicians, foremen, and industrial engineers (Hounshell 1984).

Later, when robotic arms arrived in auto plants in the 1980s for welding, painting, and lifting, predictions of massive job losses again proved premature. As output and quality standards rose, carmakers needed more design engineers, robotics technicians, quality managers, and supply-chain coordinators. Workers programming robots or maintaining automated systems often earned above-average manufacturing wages (Graetz and Michaels 2018).

Famously, while bank tellers faced automation pressure as ATMs substituted for some routine transactions, more bank teller jobs were ultimately created than destroyed: the number of tellers per branch fell but banks expanded branches and shifted teller duties toward customer-facing service tasks (Bessen 2015; Autor 2015; Klein 2019). ATMs made it economical for banks to operate more branches and expand services. While each branch needed fewer traditional tellers, the growing branch network created new roles in relationship banking, financial product sales, IT, security, and compliance.

These cases demonstrate how automation can trigger virtuous cycles: lower production costs expand markets, creating demand for new types of skilled labor that complement the technology. Chapter 11 formalizes this insight.

MARKET FAILURES MUST BE SOLVED FOR GPTs TO REACH THEIR POTENTIAL

While industrial revolutions motored by the emergence and diffusion of GPTs represent transformative periods in economic history, they do not materialize automatically through market forces alone. Significant market failures at various stages of innovation mean that GPTs are underprovided or fail to diffuse. This necessitates strategic government intervention.

First consider Romer's (1986) endogenous growth model, which is predicated on the notion that non-rival ideas generate increasing returns at the aggregate level. Unlike standard economic thinking, where diminishing returns typically curb long-term growth, knowledge spillovers enable the productivity benefits of R&D to multiply across the entire economy.

Empirical studies lend support to this perspective. Researchers have found that the social return to R&D often exceeds the private return (Griliches 1992; Hall 1996). Nordhaus (2004) estimates that innovators can capture only about 2.2% of the total surplus from innovation, with the remaining benefits accruing to society at large. Faced with such a meager slice of the bigger

pie that's created, rational private actors will consistently underinvest in the basic research required to launch a GPT.

Figure 2.2 Fewer Ideas as a Market Failure.

[INSERT FIGURE 2.2 HERE]

Notes: Supply curves are drawn as linear and parallel for expositional clarity; the qualitative results hold under more general functional forms. Q_1 = market equilibrium; Q_2 = socially optimal equilibrium. The framework follows the standard Pigouvian treatment of positive externalities (Pigou 1920).

Following Pigou (1920), Figure 2.2 explicates the logic of how this divergence between private and social returns leads to underinvestment in innovation—and how government intervention solves this market failure. The solid downward-sloping line represents the aggregate Demand for Ideas. The higher, dashed line represents Private Marginal Cost—the full, unsubsidized cost that private firms must bear to generate new ideas. The lower, dash-dot line represents the Social Marginal Cost, reflecting the "effective" cost to society when positive externalities are accounted for.

Left to its own devices, the market settles at Q_1, where firms produce fewer ideas at a higher price (P_1). The hatched triangle represents deadweight welfare loss, the economic inefficiency that occurs when the market produces fewer ideas than socially optimal. With the right government interventions, the market moves toward the socially optimal equilibrium at Q_2. By subsidizing R&D, the state effectively shifts the supply curve down to the Social Marginal Cost line, eliminating the deadweight loss.

There are four critical areas where government action can overcome these market failures: (1) undertaking or subsidizing basic science research; (2) establishing intellectual

property rights (IPRs) to incentivize development; (3) supporting commercialization; and (4) accelerating mass adoption by investing in standardization, education, and infrastructure.

Subsidizing Basic Science Research

If innovators capture only a meager slice of the value from applied R&D, the problem is exponentially worse for basic science. As Nelson (1959) argues, "basic research" is essentially the production of pure information. Since this information possesses the characteristics of a public good—it is non-rival and non-excludable—it defies standard market mechanics even more than commercial innovation does. As Arrow (1962) famously demonstrated, private entities systematically underinvest here because they cannot fully appropriate the benefits.

Figure 2.3 Different Types of Goods

[INSERT FIGURE 2.3 HERE]

Notes: Excludability refers to the ability to prevent others from consuming a good. Rivalry refers to whether one person's consumption diminishes the quantity available for others. Private goods exhibit both properties; public goods exhibit neither. Knowledge and ideas—the outputs of basic research—are classic public goods: once produced, they can be consumed by additional users at zero marginal cost (non-rivalrous), and it is difficult to prevent others from accessing them (non-excludable). This combination creates the appropriability problem central to the economics of innovation (Nelson 1959; Arrow 1962).

Figure 2.3 compares different types of goods based on two fundamental characteristics: excludability and rivalry. Private goods (top-left) exhibit both properties and function well in markets. Common-pool resources (top-right) are rivalrous but difficult to exclude, making them susceptible to the "tragedy of the commons." Club goods (bottom-left) can be controlled through

access restrictions while remaining non rivalrous. Public goods (bottom-right)—including basic scientific research, national defense, and street lighting—are neither excludable nor rivalrous.

A free-rider problem bedevils the provision of public goods such as basic scientific research, and thus the marketplace underprovides it. This reflects a classic Prisoner's Dilemma: even if all firms would be better off collectively funding basic research, the dominant strategy for any individual firm is to defect. Without a binding mechanism to enforce cooperation, the research is never funded.

Governments therefore play crucial roles in funding public research institutions and subsidizing universities. The US Federal Government helped fund much of the basic research that underwrote the Third Industrial Revolution. Starting in World War II, the government directed substantial public funding to leading universities and research labs. During the Cold War, Washington financed basic and applied scientific research through DARPA, the NSF, the Atomic Energy Commission, the Office of Naval Research, and the NIH. These efforts significantly accelerated development of key technologies including semiconductors, GPS, and the internet (see Mazzucato 2013).

Intellectual Property Rights

Yet another type of market failure that can impede the emergence and diffusion of GPTs, and ultimately the birth of industrial revolutions, is a lack of clarity over property rights. Even if would-be inventors are armed with the best of ideas for developing and exploiting commercially ripe new technologies, private investment can stall when inventors, investors, and entrepreneurs cannot appropriate a return on their discoveries and ideas (Arrow 1962).

Producing the ideas is only half the problem. Without enforceable property rights, knowledge can be copied the moment it becomes public, leaving innovators no way to recoup

their investment. The government must therefore also define and enforce the rights that let innovators capture enough surplus to justify commercialization.

Patents achieve this by shifting ideas from pure public goods into something closer to club goods: the underlying knowledge remains non-rival, but legal excludability lets inventors appropriate a return. Figure 2.3's bottom left-hand quadrant conveys this: ideas approach something akin to a club good, unlike the public good of basic scientific knowledge that is not excludable.

Importantly, this transformation works less by conferring monopoly power—rivals can typically "invent around" disclosed claims (see Haber 2016)—than by enabling a market for ideas in which divisible, tradeable rights allow inventors to specialize in discovery while other parties specialize in commercialization, manufacturing, or distribution (Spulber 2015).

Licensing contracts convey the tacit knowledge that formal disclosure cannot capture—in the form of blueprints, onsite training, and process adjustments that translate patented concepts into working products. AT&T's transistor licensing program exemplifies this logic; the Bell Labs symposia of the early 1950s bundled patent access with hands-on manufacturing know-how, rapidly diffusing semiconductor technology across dozens of firms (Nagler et al. 2022).

Yet this system faces a vulnerability: the "hold-up" problem. An investor may refuse to make a relationship-specific investment if she fears that, once the investment is sunk, her counterparty can renegotiate or expropriate the surplus (Klein et al. 1978; Williamson 1985).

Patent licensing contracts can be written (and enforced) to mitigate hold-up by diversifying supply relationships and committing parties to terms before investments are sunk. When IBM developed the original PC, released in 1981, it insisted on at least two microprocessor suppliers to ensure supply chain stability—effectively compelling Intel to license

its technology to AMD as a "second source" guarantee. This helped cement the x86 architecture as the dominant industry standard (Grindley and Teece 1997). In Chapter 6 I will explain how upstream chip designers protected themselves from downstream opportunism in similar ways during the smartphone era.

By transforming proprietary inventions into industry-wide standards, firms like Texas Instruments and Intel positioned themselves at the center of a vast web of technological interdependence. As I will show in Chapter 5, these "superstar" firms evolved into central hubs of an "innovation commons." These patents became "standard essential"—legally required for any firm implementing the standard—raising new questions about licensing terms that Chapters 3 and 6 will address.

In short, patents commercialize ideas not by conferring monopoly power but by creating tradeable assets, enabling specialization, and bundling tacit knowledge with formal rights.

Other Government Roles

Market mechanisms alone often fail to achieve optimal technological standardization due to coordination problems and competing private interests (Farrell and Saloner 1985). Governments play essential roles in overcoming these market failures.

First, they can act as a coordinator or regulator, explicitly mandating a unified standard or convening standard-setting bodies. Consider the government-led standardization of railroad gauges in the 19th Century and electric power systems in the early 20th Century (Shapiro and Varian 1999). In the 1980s, the US federal government resolved the "protocol war" between TCP/IP and the international OSI standard by mandating adoption of TCP/IP across its military and research networks, effectively establishing a dominant design (Abbate 1999).

Second, by providing massive, guaranteed demand as a "customer of first resort," the government can make a specific design the salient choice. Consider that the US military and NASA were the primary buyers of early transistors, purchasing them in bulk—and often at premium prices (Misa 1985). Texas Instruments secured major defense work for its first generation of integrated circuits, including winning the bid for the Minuteman II missile guidance system (Texas Instruments 2021).

Finally, there is the government as a nudger. Agencies like NIST facilitate standardization processes that enable compatibility and interoperability. Rather than impose standards coercively, they serve as a neutral arbiter, convening industry stakeholders to develop voluntary consensus standards (Tassey 2000). The rise of voluntary Standards-Setting Organizations (SSOs) embodies this role. In the next chapter, I will show how the US federal government's subtle, yet decisive involvement in helping private actors develop wireless telecommunications standards helped underpin the digital revolution.

Yet coordination is only half the problem. Even when firms agree that a standard is needed, they fight over whose technology becomes the standard—and thus who captures the largest share of the innovation rents. This "Battle of the Sexes" dynamic means governments and SSOs must manage not just coordination, but distribution, designing rules (like FRAND commitments) that stabilize bargains and prevent holdouts from derailing collective adoption.

Governments also help diffuse technologies once standards are set. SEMATECH, a consortium established in 1987 by the federal government and fourteen leading US semiconductor firms, pooled resources to standardize manufacturing processes and reverse market-share losses to Japan (Irwin and Klenow 1996). The Manufacturing Extension Partnership provided technical assistance to help local manufacturers adopt Computer-Aided

Design and Manufacturing software (Shapira 2001). These interventions worked not by picking winners but by breaking coordination traps that left firms waiting for others to move first.

Beyond facilitating technological standardization and diffusion, governments address other market failures that private actors cannot solve alone. Because infrastructure and skilled labor are themselves public goods, firms underinvest in both; the Tennessee Valley Authority's electrification program, for instance, lowered input costs for heavy industry and ignited a sustained private employment boom (Kline and Moretti 2014). The SBIR (Small Business Innovation Research) program bridged the "Valley of Death" between concept and commercialization for upstart firms like Qualcomm, helping pave the way for the smartphone era (Lerner 1999). The Morrill Land-Grant Acts and the G.I. Bill built the human capital base for successive industrial revolutions that involved these and other technologies.

CONCLUSION

Industrial revolutions hinge on (1) standardization, which ensures interoperability and economies of scale; (2) organizational adaptation, through which firms retool operations and retrain workforces; and (3) government support that solves multiple market failures across several domains. That includes providing strong patent protections for ideas and financing for early-stage innovation. It also includes basic research funding, infrastructure buildup, education and training, and help with technological standardization and diffusion efforts.

The lessons from the past three industrial revolutions set the stage for the rest of this book. Capturing AI's full potential will demand not only scientific progress and entrepreneurial ingenuity, but also the slow, unglamorous work of standardizing tools and interfaces, financing complementary investments, and rebuilding organizations so that frontier capability becomes routine in day-to-day production. These are topics I take up in Chapters 10 and 11.

Each industrial revolution also encountered a "productivity paradox": measured output often sputters before a new technology's benefits can be fully realized. Companies initially pour resources into learning, reorganizing, and complementary investments—which can dampen productivity metrics—yet eventually, once the reorganizations are complete, technologies diffuse broadly, bringing exponential cost reductions and long-delayed but dramatic gains in performance and affordability. Whether steam, electricity, or the microprocessor, once the core GPT is standardized and widely adopted, subsequent improvements tend to follow rapid or exponential curves, fueling fresh applications, surging demand, and new industries.

In Chapter 11, I will explore this phenomenon alongside other macroeconomic and socio-economic shifts, including wages, jobs, profits, and distribution. There, I formalize the key historical mechanism from this chapter: the gap between what a frontier GPT can do and what the economy can realize before standards, complements, and organizations catch up. I will inform these modeling exercises with the historical evidence reviewed here: that, while automation can displace specific tasks, growing markets and new skill demands often increase overall employment.

However, before we get there, the next few chapters move from the general historical patterns of the past to the specific origins of the Artificial Intelligence revolution. Chapter 3 begins this investigation by exploring the unique policy environment that birthed the Third Industrial Revolution and set the stage for the Fourth. I explore how, between the late 1970s and 2000s, a broad "populist-statist" consensus was supplanted by a new paradigm. Earning strong bipartisan support, this new approach prioritized evidence-based regulation, established novel property rights over data and software, and aggressively reduced transaction costs. This specific

institutional mix solved the market failures described above in a new way, eventually laying the legal and financial groundwork for the AI boom.

CHAPTER 3: THE CREATIVE DESTRUCTION PARADIGM

Over the late 20th century, US policy changed drastically. Early-era "populist-statist" leaders piled on price controls, trade barriers, and make-work schemes to shore up political constituencies—often at the expense of innovation. Gradually, however, a new paradigm dedicated to promoting innovation and technological diffusion, which I call the **Creative Destruction Paradigm (CDP)**, took hold. It internalized Joseph Schumpeter's concept of "creative destruction"—the process by which profit-seeking firms innovate, disrupt, and ultimately displace vintage industries and incumbents to make way for new technologies and goods and services (see Aghion et al. 2014). Over a few short decades, it appreciably reduced transaction costs, solved coordination failures, and managed distributional conflicts in the service of technological progress.

Between the Carter and Obama presidencies, the CDP's architects built an evidence-based, cost-benefit framework that transformed the federal government's approach to trade policy, IP, telecommunications, electronic contracting, data governance, and competition policy. Quietly, but steadily, this approach grew to enjoy wide support. US policymakers of all stripes came to believe that, while markets should generally guide the commercialization of innovation, the government provided an essential helping hand by clarifying rules of the road and furnishing the right incentive structure. For example, Democrats and Republicans alike celebrated the passage of Section 230 of the 1996 Telecommunications Decency Act as instrumental for securing the internet's freewheeling commercialization after immunizing digital platforms from legal liability for harm caused by third parties (Kosseff 2019).

In this chapter, I trace that arc, showing how both parties came to embrace the CDP in doctrine, statute, and implementation, and explain why this convergence—not mere "neoliberal"

deregulation or the so-called Washington Consensus (Williamson 1990)—set the stage for the digital economy and ensuing AI revolution. I argue that, unlike the CDP, neither Neoliberalism nor the Washington Consensus ever obtained widespread support within the US among politicians or the public. The latter's emphasis on market liberalization, deregulation, and fiscal austerity—more on this below—often met with fierce resistance, and even their proponents later acknowledged that these agendas never really enjoyed a consensus, neither in the American capital nor anywhere else (see Spence 2021; Menand 2023). More importantly, neither Neoliberalism nor the Washington Consensus were ever truly implemented in the US; instead, it was the CDP that dominated the period spanning from Carter to Trump.

However, the CDP was never about state-directed industrial policy either. As I will demonstrate in Chapters 4-8, it was the **private sector** that ultimately commercialized innovation during the Third Industrial Revolution's latter stage, coinciding with the CDP era. Profit-seeking firms introduced and diffused new goods and services that included cloud computing networks, smartphones, digital platforms, and the gig economy.

Yet, while this market-based commercialization engine was neither centralized nor directed by the state, it depended on the government's CDP-inspired rules and regulations to help solve complex collective action problems that otherwise prevent firms from jointly finding ways to establish interoperability, construct new supply chains, and bring new products to market. It took the US federal government's concerted, multi-decade effort to erect the necessary framework across IP, technology transfer, and telecommunications to solve several of the market failures I outlined in the previous chapter.

Consider two examples. First, the Bayh-Dole Act allowed universities and small businesses to own patents on inventions resulting from federally funded research, spurring

innovation in biotechnology and computing by incentivizing commercialization (US Congress 1980b). Second, the Digital Millennium Copyright Act (DMCA) Safe Harbor provisions introduced protections for online service providers against copyright infringement claims for user-generated content (US Congress 1998).

During the CDP era, the federal government also encouraged, if not cultivated, the formation of private, voluntary organizations that brought together various players operating in the digital device supply chain, including potential rivals. These included standard-setting organizations (SSOs) that establish technical standards to ensure compatibility and interoperability among products and technologies. By determining which patented technologies are incorporated into new products, SSOs help unify the industry around common platforms. As I shall demonstrate in greater detail further below, and in Chapter 6 as well, SSOs played a pivotal role in standardizing mobile telecommunications technologies and were critical to the development and commercialization of smartphones and similar digital devices.

The invention and commercialization of new technologies during the CDP era was also a byproduct of the federal government's deregulation of industries previously governed by strict regulations. For example, the 1996 Telecommunications Act overhauled the telecommunications industry by reducing barriers to entry and incentivizing and enabling new companies to challenge established providers (US Congress 1996). This legislation prompted both incumbents and newcomers to expand, improve, and accelerate the deployment of their communications networks, whether broadband, wireless, or mobile. These efforts were further facilitated by multiple spectrum auctions conducted by the FCC that efficiently allocated wireless frequencies to the network providers best suited to supply the market. This, in turn, encouraged significant growth and innovation in internet services and cellular coverage (see Chapter 8).

Therefore, while I share Mazzucato's rejection of the minimalist, neoliberal view of the state—acknowledging, as she does, that public investments (e.g., DARPA, NIH) were foundational to the IT revolution—our diagnoses of the mechanism differ. Mazzucato (2013) argues for an "entrepreneurial state" that actively directs innovation, takes on commercial risk, and creates markets through mission-oriented policies. By contrast, the CDP posits that the government's primary role was **institutional scaffolding**: solving the collective action, coordination, and commitment problems (see Chapter 2) that prevented private firms from investing. In my framework, the state did not "direct" the commercialization of the smartphone or the digital economy; rather, it established the rules, standards, and property rights that allowed the private sector to successfully commercialize and diffuse these technologies.

To appreciate the scale of this transformation, it is worth contrasting the CDP with the regime it replaced. While I will explore the pre-CDP era in greater detail further below, some stark examples of **populist-statist policies** help contextualize the shift. Early-20th century Americans saw farm price supports guarantee favorable minimum rates to small landholders, shielding them from volatile markets but encouraging overproduction that simultaneously increased food prices and ate up scarce public revenues. Similarly, mid-century wage-and-price freezes (e.g., Nixon's 1971 controls) placated labor constituencies but disrupted normal supply-and-demand signals, fostering shortages and black markets. When President Trump began floating price controls of various sorts—including strict caps on credit card interest rates and pharmaceutical prices—during his second term, the specter of these distortions returned to haunt Washington, D.C.

This chapter continues as follows. First, I trace the historical evolution from the populist-statist consensus of the early 20th century to the emergence of the CDP, detailing how cost-

benefit analysis and evidence-based rulemaking became institutionalized in the Executive Branch. Second, I examine the interstitial judicial shifts—including key Supreme Court decisions on patentability—that quietly laid the groundwork for broader reform. Third, I explore the five structural pillars of the CDP outlined in Table 3.2 below: the centralization of patent appeals under the CAFC; the validation of software patents; the Bayh-Dole Act's university technology transfer regime; the globalization of IP standards via the WTO; and the governance of Standard Essential Patents (SEPs). Finally, I analyze how the CDP was transaction costs reducing across the digital economy.

THE CDP VS. SOME TIRED SHIBBOLETHS

The key technological advancements that this book seeks to explain came about as the result of profound reforms to IPRs, trade policy, telecommunications, antitrust regulations, and a host of other regulatory domains between the late 1970s and the early 2000s. However, these changes are not best understood as embodying so-called Neoliberalism or the "Washington Consensus."

Neoliberalism is an ideology that emerged prominently in the late 20th Century, advocating for free-market capitalism as the primary means to achieve economic growth. It emphasizes the static efficiency of market mechanisms in allocating resources and reducing the state's role via privatization, deregulation, and globalization (see Harvey 2005; Mirowski and Plehwe 2009). Its central tenet is that the government is a minimalist referee: any intervention beyond maintaining property rights or enforcing contracts risks distorting prices.

Similarly, the Washington Consensus refers to a specific set of policy prescriptions formulated by economist John Williamson (see Williamson 1990). Promoted by institutions like the IMF and World Bank (Stiglitz 2006), these policies were designed primarily for developing countries facing macroeconomic crises. They emphasized fiscal discipline, trade liberalization,

and the privatization of state enterprises. The overarching goal was stabilization: fixing economies that suffered from fiscal imbalances, rather than explicitly engineering the frontier of technological growth.

This book and especially this chapter challenge the view that these ideologies birthed today's globalized knowledge economy. Instead, I attribute the groundwork for the AI revolution to a distinctly bipartisan agenda that prioritized dynamic innovation through evidence-based policy. Over several decades, a series of unheralded laws and regulations shifted the US approach from a populist-statist focus to a pragmatic, technocratic emphasis on innovation. Under the CDP, the government played a decisive, complementary role—investing in R&D, setting standards, and creating markets—rather than simply stepping back and hoping for the best.

Table 3.1 outlines the key differences between these paradigms. It clarifies why the CDP, with its willingness to tolerate short-term disruption to catalyze long-run productivity, set the stage for the digital revolution, whereas the static focus of Neoliberalism and the stabilization focus of the Washington Consensus did not.[3]

Table 3.1 Comparing the CDP to Neo-liberalism and the Washington Consensus

[3] Of course, Neoliberal architects (including the Chicago School) believed that deregulation and open trade would spur growth. Yet, their primary analytical lens was static allocative efficiency, not the Schumpeterian creative destruction central to the CDP. While they assumed innovation would naturally follow market liberalization, the CDP proactively engineers it. I revisit this distinction in Chapter 7.

	Creative-Destruction Paradigm (CDP)	*Neoliberalism*	*Washington Consensus*
Efficiency Focus	**Dynamic** – prioritizes long-run productivity growth via continual innovation and market entry/exit cycles.	**Static** – emphasizes allocative efficiency today, minimizing rents and deadweight loss in existing markets.	**Static** – stresses short-run stabilization and allocative gains from liberalization and privatization.
Role of Government	**Active, adaptive, muscular** – designs and updates rules to foster experimentation, correct market failures, co-invests in frontier technologies.	**Minimalist, referee-only** – government's job is to set "rules of the game" and otherwise step back.	**Enforcer of orthodoxy** – government implements externally prescribed macro and micro reforms; once "fixed," moves toward hands-off.

Outcome Horizon	**Long-run growth** – tolerates short-term disruption to reap future productivity gains.	**Short-run cost savings** – seeks immediate price and fiscal benefits.	**Crisis exit & stabilization** – first secure macro balance, then hope market forces drive recovery.

Notes: This table contrasts three major policy paradigms—Creative-Destruction (CDP), Neoliberalism, and the Washington Consensus. The CDP emphasizes long-run, innovation-driven growth; Neoliberalism centers on near-term allocative efficiency; the Washington Consensus focuses on macro-stability/market liberalization in crisis-hit economies.

These differences extend to policy domains and tools. Under the CDP, governments may have deployed targeted R&D grants, engaged in forward-looking antitrust enforcement, helped standardize nascent technologies, and endorsed public-private partnerships. Neoliberal prescriptions, by contrast, favor broad deregulation and labor-market flexibility, assuming markets alone will guide investment optimally. The Washington Consensus adds conditional lending and structural adjustment templates to stabilize economies in crisis.[4]

To be sure, superimposed upon the CDP's concerted efforts to address the type of market failures outlined in the previous chapter, and which the rest of this chapter will explore in greater detail, Washington, D.C. also adopted many policies aligned with Neoliberalism: taming

[4] The classic Washington Consensus was originally targeted at emerging-market crises. Nevertheless, its ethos of fiscal restraint and liberalization frequently influenced domestic U.S. policy, particularly in the 1990s (see Stiglitz 2006).

inflation; dismantling regulatory barriers; privatizing some industries, cutting taxes that distorted investment decisions, and prizing labor deregulation (see Niskanen 1988). Moreover, as the US government pushed other countries to liberalize trade and financial flows, it viewed orthodox macroeconomic stabilization measures associated with the Washington Consensus as precursors to reforms. In the 1990s and 2000s, following debt crises and emerging-market collapses, Washington, D.C. and the IMF promoted policies of fiscal restraint, if not outright austerity, in fragile economies, coupled with calls for deregulation and the elimination of price controls and other economic distortions (see Stiglitz 2002).[5]

However, the main point stands: the US federal government often spoke out of both sides of its mouth in the runup to the new millennium, lionizing the virtues of fiscal restraint and macroeconomic orthodoxy while acting as the strategic conductor for a diverse ensemble of universities, private firms, government agencies, and investors, ensuring they played in harmony even when their short-run interests were misaligned. And doing this often took precious money, time, and attention, not disinterested aloofness. As the previous chapter made clear, seeking immediate cost savings or macroeconomic stability, rather than focusing on actively nurturing tomorrow's technologies by solving the market failures that impede their creation and diffusion, will not get you tomorrow's technologies.

Before detailing the specific policies that underpinned the digital economy—and eventually the AI revolution—allow me to briefly review the US presidential administrations that contributed to this phenomenon.

---

[5] Note, however, that its succeeding paradigm, the so-called Post-Washington Consensus, did include stronger governance and social-safety-net elements (Rodrik 2006).

Cursory View of Some Major CDP Reforms

By reshaping the American political economy in subtle ways, President Jimmy Carter (1977-1981), played a pivotal, often underappreciated, role in belatedly ushering in the digital revolution. Rather than throwing his weight behind a single, headline-grabbing statute, his legacy lies in a constellation of deregulatory and pro-innovation reforms whose aggregate effect was to slash costs, open markets, and unleash new sources of technological development.

While several major regulatory reforms in the late 1970s and early 1980s were not necessarily about innovation per se, they laid the groundwork for a more dynamic, service-oriented economy (see Chapter 4). The Airline Deregulation Act of 1978, for example, phased out federal route and fare controls, unleashing competition that led to a significant decline in airfares—studies estimate a real drop of about 20–33% over the following 15 years—and a substantial increase in passenger traffic (Winston and Morrison 1997). The Staggers Rail Act of 1980 freed railroads from many rate and service mandates, resulting in a roughly 40% decline in inflation-adjusted rail rates over the subsequent decades (AAR 2025). The Motor Carrier Act of 1980 eased entry and pricing rules for trucking firms, leading to lower shipping rates and increased competition (Paxson 1981).

In the energy sector, the Public Utilities Regulatory Policies Act of 1978 (PURPA) required utilities to purchase power from qualifying small producers, catalyzing the independent power sector (US Congress 1978; EIA 2018). Finally, the National Energy Act of 1978, through the Natural Gas Policy Act, began the process of deregulating wellhead natural gas prices, helping to align supply with market demand (NaturalGas.org 2013).

Collectively, these measures contributed to a dramatic reduction in the logistical cost-share of US GDP, which fell from about 16% in the early 1980s to roughly 10% by the late

1990s (FHWA 2002). This transformation in logistics costs and reliability helped pave the way for just-in-time supply chains, geographically dispersed production networks, and the high-velocity data flows that manage them, a topic I will revisit in ensuing chapters.

Carter also spearheaded regulations that bore a more direct impact on innovation, allowing the US to take the pole position in high-tech sectors even as its postwar manufacturing dominance waned (see Carter 1980a). He sought to phase out price controls and eliminate unnecessary, duplicative, and conflicting regulations across federal agencies, both of which hindered technological commercialization (see US Congress 1979a). Carter also pursued efforts to translate federally funded inventions into marketable IP, a topic I will explore in greater detail ahead.

Often, Carter's CDP approach took the form of championing what are unheralded, yet crucially important, bills. For example, as I will explain shortly, the Stevenson-Wydler Technology Innovation Act (see US Congress 1980a) established federal lab technology-transfer offices and codified the National Technical Information Service's (NTIS) role in disseminating basic research (see Schacht and Thomas 2006). Furthermore, as Chapter 7 will show, under Carter US competition policy quietly began to shift to promoting innovation, a focus that outlived his administration and, as Chapter 7 will show, allowed digital platforms to grow organically and via acquisitions, establishing ecosystems of bundled, highly integrated services that maximized user engagement and helped pave the way for the AI Revolution (see Chapter 8).

Reagan also advanced policies that reduced government intervention and promoted competition (see Niskanen 1988) in ways that redounded positively for innovation. While this included significant tax reforms that lowered marginal income tax rates and simplified the tax code, as well as financial sector deregulation (see O'Mara 2019), he also built upon Carter's

efforts to strengthen IPRs—notably through the establishment of the Court of Appeals for the Federal Circuit in 1982, a topic I will explore ahead—and consolidated a more technocratic approach to antitrust (see Chapter 7).

President Clinton, especially, embraced technological advancement. While his administration's overhaul of telecommunications through the 1996 Act provided the physical foundation for the digital economy (a topic I return to in Chapter 8), his CDP legacy also includes promoting the rise of the internet more directly through the "Framework for Global Electronic Commerce," which advocated for the private sector's lead in developing the internet (Clinton and Gore 1997), and the DMCA, which updated copyright law for the digital age (US Congress 1998).

Finally, building on Carter's foundation, Presidents Ronald Reagan, George H.W. Bush, and Bill Clinton all advanced freer trade. While President George H.W. Bush initiated the negotiations for the North American Free Trade Agreement (NAFTA), it was Clinton who secured its passage through Congress in 1994. He also played a pivotal role in establishing the World Trade Organization (WTO) in 1995 and worked to grant Permanent Normal Trade Relations status to China, facilitating its WTO entry in 2001. These agreements lowered trade barriers, encouraged the expansion of global, vertically disintegrated supply chains, and allowed US companies to access international markets more effectively—all topics I will explore in the next chapter.

A TECHNOCRATIC APPROACH DISPLACED A POPULIST ONE

Before discussing the transition to the CDP and how it differed from the earlier political and regulatory paradigm, I will explain its predecessor, which I refer to as the populist-statist paradigm, and review its checkered history. The lessons from this prior approach are that

prioritizing short-term political redistribution over market efficiency tends to entrench incumbents, distort investment signals, and ultimately stifle the dynamic competition necessary for technological progress.

The Populist-Statist Paradigm

Populist policymakers deploy a toolbox of distortionary interventions designed less to correct market failures than to cement political alliances through rent creation and redistribution. At its core, populism trades long-run growth for immediate, targeted gains, using a variety of devices that generate excludable benefits for key constituencies. Indeed, populism is paradoxical in this way: while often rhetorically attacking "elites," populists exploit executive and legislative levers to disadvantage diffuse entrepreneurs, laborers, and consumers in favor of concentrated interests. They therefore build political loyalty amongst protected incumbents and workers at the expense of market entry, greater competition, and the dynamic gains born of creative destruction.[6]

Consider a set of populist strategies with mostly domestic ramifications. First, price-distorting measures—such as direct controls on food, fuel, or utilities—artificially cap consumer prices or prop up producer incomes. Second, credit and financial repression steers subsidized loans and interest-rate caps to politically connected firms while throttling market-rate borrowing for entrepreneurs. Third, state hires—public-sector jobs distributed as patronage—create a dependable vote-bank at the cost of bloated payrolls and stagnant productivity.

There are also fiscal giveaways and targeted subsidies that deserve their own attention. By spending lavishly on make-work programs and white elephants, populists funnel government spending into politically powerful regions or demographics that swing elections. While those

[6] This sequence of paragraphs draws on Magistro and Menaldo (2025).

interventions may temporarily lift incomes, they tend to crowd out private investment and tie up scarce public resources, undermining the infrastructure and R&D spending necessary for technological progress (see Magistro and Menaldo 2025).

Finally, populist strategies extend to the international arena through trade protectionism. Tariffs, import surcharges, and voluntary export restraints are rolled out to defend domestic industries viewed as electoral bulwarks. By raising the cost of foreign inputs or finished goods, populists generate rents for sheltered firms and preserve local jobs, yet they curb competition, raise prices, provoke retaliatory barriers that hurt exporters, and choke off innovation (ibid).

The longstanding populist-statist consensus that spanned much of the 20th Century condoned significant government intervention to address social, economic, and environmental problems. However, it often prioritized aggressive regulatory action over the empirical assessment of policies' effectiveness—instead taking on a moral dimension articulated in self-evident terms such as protecting "the little guy" from the "corrupt establishment." This ironic contradiction culminated in political capture and regulatory overreach that introduced economic distortions, encouraged further rent-seeking, and dampened innovation (see Buchanan and Tullock 1962; Menaldo and Wittstock 2025; Peltzman 1976; Stigler 1971).

First Stage of American Populist-Statism

The first stage of the populist-statist era began between the 1890s and 1920s. It was a period that saw a significant, coherent expansion of the government's intervention in the US economy. Presidents Theodore Roosevelt and Woodrow Wilson championed so-called progressive policies that increased federal oversight to ostensibly address several challenges engendered by

industrialization, urbanization, and increased economic inequality and championed the "common man and woman" over "the rich and powerful."[7]

During this early stage, two policies adopted in 1906 stand out. The first was the Hepburn Act, which expanded the Interstate Commerce Commission's power to set railroad rates and was explicitly pitched as protecting small shippers and farmers who were crushed by steadily rising "long-haul" rates. The rate controls also aimed to limit railroad earnings to a "fair return" of about 6 percent, paving the way for several carriers to eventually exit the market (Scribner 2013). The second was Meat Inspection Act, which inaugurated federal policing of food safety and were framed as a crusade to protect the "little guy" from unscrupulous packers. Because they required durable investments in sanitary facilities, process designs, and continuous documentation that were difficult for small firms to incur, the industry underwent a massive, albeit a belated, shakeout: the number of smaller, state-inspected plants plummeted by 30%

[7] Famously, these policies included aggressive antitrust actions against big, profitable companies—but not necessarily monopolies in the technical sense of the term. Instead, policymakers punished bigness and success by sometimes breaking up large industrial trusts, even if their size or their industry's degree of concentration did not adversely affect prices or technological progress. While in Chapter 7 I revisit populist competition policy in the US context, in this section I focus exclusively on other historical examples of populist regulation.

while federally inspected facilities expanded by 50% between 1972 and 1976 (Ollinger and Mueller 2003).[8]

Similarly, during the 1910s and early 1920s, many states expanded occupational-licensing requirements—adding exams, fees, and residency rules for barbers, beauticians, real-estate agents, and others. While marketed as consumer safeguards, these licensing regimes generated rents for incumbent practitioners by raising barriers to entry and insulating them from new competitors. This meant less competition and higher prices for those very consumers these regulations were ostensibly intended to help (see Stigler 1971).

Similarly, during World War I, Wilson adopted price controls on food and fuel and established the National War Labor Board (NWLB) to arbitrate labor disputes and stabilize industrial production. The Board set uniform wage increases, approved cost-of-living adjustments, and enforced no-strike pledges in essential industries (see Rockoff 1984). Though framed as a patriotic necessity, these binding wage directives delivered targeted gains to unionized workers at the expense of those who were not unionized; not only did it strengthen labor's political alliance with the Wilson administration, but it disadvantaged non-unionized workers by pricing them out of labor markets (Vedder and Gallaway 1993).

For his part, President Herbert Hoover signed the Smoot–Hawley Tariff Act in 1930, ramping up protectionism. Steep increases in import duties across hundreds of product lines were designed to shelter U.S. farmers and manufacturers. However, this came at the expense of lost

[8] The Wholesome Meat Act of 1967 and the Wholesome Poultry Products Act of 1968 expanded the scope of inspections and sanitation standards, further increasing the regulatory burden on smaller plants (Ollinger and Mueller 2003).

export markets; it also triggered retaliatory tariffs that deepened the Great Depression (Irwin 1996).

By introducing extensive federal programs in response to the Great Depression during the 1930s, FDR's New Deal further entrenched the populist-statist approach. The National Industrial Recovery Act attempted to manage industrial production and pricing through codes of so-called fair competition, while the Agricultural Adjustment Act introduced direct price-floor subsidies and acreage controls to prop up farm incomes at the expense of consumers (see Fishback 2007). The National Labor Relations Act (Wagner Act) created a statutory right to unionize and bargain collectively and was framed as defending ordinary workers against corporate power; but it created barriers to entry to younger laborers from entering the workforce and redistributed from consumers and non-unionized workers to unionized laborers (see Vedder and Gallaway 1993).[9] Regulation Q, adopted in 1933, capped the interest rates that banks could pay on deposits, reducing competitive pressures and forcing many savers—and the funds they might have invested in emerging businesses—into lower-yield, government-insured accounts.

Finally, to absorb mass unemployment and shore up political support, the Roosevelt administration launched public employment programs on an unprecedented scale. The Works Progress Administration (WPA) and the Civilian Conservation Corps (CCC) provided government-funded jobs to millions of Americans in projects ranging from road building to the

[9] Any hit to consumers came about secondarily, via cost-pass-through in goods prices, and varied greatly by industry and time. Most redistribution was directly from firms' profit margins to unionized employees (see Epstein 2006).

arts. These positions provided immediate, targeted relief—and durable voter loyalty—even as they swelled the federal payroll and potentially diverted labor from the private economy.[10]

Second Stage of Populism-Statism

During the 1960s, President Lyndon Johnson amplified the statist-populist approach outlined above. Overcoming fierce industry opposition, his administration signed the Fair Labor Standards Amendments of 1966, which expanded federal labor standards to an estimated 9.1 million previously uncovered workers in sectors like agriculture, retail, and hospitality (DoL 2024). The amendments raised the statutory federal minimum wage in stages, to $1.60 by 1968, the historic peak of the federal minimum wage's purchasing power (ibid). While these increases significantly raised wages for low-income workers, and particularly for Black Americans (Bailey et al. 2021), the higher wage floor reduced labor force participation or employment among the most vulnerable groups, such as teenagers and low-skilled workers (Wessels 2001; Neumark and Wascher 2008).

Similarly, through the Food and Agriculture Act of 1965, Congress solidified a system of managed pricing by coupling price-support loans with direct income payments to farmers (US Congress 1965), ensuring that participating producers received returns often exceeding market

[10] For example, Benjamin and Matthews (1992) employed state-level panel data and discovered employment crowd-out effects during both the First and Second New Deal periods. Meanwhile, Liu and Fishback (2019) find that New Deal public works programs had only modest positive effects on private-sector earnings and no significant effect on private employment, while relief programs were associated with lower private employment and reduced occupational mobility, likely due to crowding out effects.

levels (see Cochrane 1979). These supports aimed to stabilize farm incomes and manage supply. Yet, by shielding farmers from market prices, the policy entrenched a costly subsidy structure that prioritized the preservation of the "family farm" over economic efficiency (see Gardner 2002). This induced distortion in turn sustained barriers to entry in the agricultural sector and led to higher food prices (Edwards 2005).

President Richard Nixon, LBJ's successor, continued in his predecessor's footsteps. He adopted wage and price controls between 1971 and 1974, including a 90-day freeze on all wages and prices, followed by graduated controls enforced by the so-called Cost of Living Council—an unprecedented peacetime federal intervention in private contracts (see Rockoff 1984). Nixon also introduced a temporary 10% surcharge on imports, designed to force trading partners into currency-realignment negotiations (see Irwin 2013). Finally, he unilaterally closed the gold window, ending the dollar's convertibility to gold and effectively dismantling the Bretton Woods international monetary system (see Gowa 1983). This resulted in higher inflation and hints of fiscal dominance—the Federal Reserve monetizing the US debt (see Meltzer 2005).

Populism and Relatively Weak IPRs

While Chapter 7 documents how populism extended to antitrust both before, during, and after the Nixon administration, here I briefly discuss its influence on IP policy. In 1970, informed by the idea that patent owners automatically exercise monopoly power—which, as I explained in the previous chapter, is an analytically and empirically dubious view—the DoJ declared nine specific patent licensing practices per se illegal, a "per se" approach fittingly referred to as the "Nine-No-Nos" (see Gilbert and Shapiro 1997).

Specifically, the DOJ essentially outlawed the following licensing policies: (1) royalties not reasonably related to sales of the patented products; (2) restraints on licensees' commerce

outside the scope of the patent (tie-outs); (3) requiring the licensee to purchase unpatented materials from the licensor (tie-ins); (4) mandatory package licensing; (5) requiring the licensee to assign to the patentee patents that may be issued to the licensee after the licensing arrangement is executed (exclusive grant-backs); (6) licensee veto power over grants of further licenses; (7) restraints on sales of unpatented products made with a patented process; (8) post-sale restraints on resale; and (9) setting minimum prices on resale of the patent products (ibid).

By chilling otherwise pro-competitive patent-pooling and cross-licensing arrangements, these policies drove licensing underground, increased transaction costs, and deterred the very collaboration that fuels complex technology development. Industries such as semiconductors and telecommunications, where overlapping patents are common, were especially hampered, as firms became wary of entering into necessary cooperative agreements. The increased legal uncertainty and risk of antitrust liability hampered companies' ability to efficiently allocate rights and share risks, stifling incentives to innovate (Gilbert et al. 1997).

The formal abandonment of this framework was gradual. A 1981 speech by Deputy Assistant Attorney General Abbott Lipsky first signaled the DoJ's retreat from per se rules (Lipsky 1981). Then, the 1988 Antitrust Enforcement Guidelines for International Operations moved to rule-of-reason analysis—one based on cost-benefit analysis (DoJ 1988). Finally, the 1995 DoJ-FTC Antitrust Guidelines for the Licensing of Intellectual Property definitively buried the framework (DoJ and FTC 1995). In Chapter 7 I will outline the general transition in US competition policy to a rule-of-reason approach, and away from a per se one.

Why the CDP Approach is So Exceptional

The CDP represented a rejection of the politically transactional, redistributive logic that spanned from FDR to Nixon. In the face of stagflation—the pernicious mix of inflation and economic

recession—and worries about American competitiveness, the statist-populist model imploded (Gerstle 2022), creating space for a bipartisan coalition of technocrats who eventually replaced political patronage with cost-benefit analysis (CBA), substituted price controls with competitive markets, prioritized the diffuse benefits of consumer welfare over the concentrated rents of protected industries, and elevated innovation to the preeminent economic concern.[11]

But the CDP went beyond undoing the damage caused by populism. As Chapter 2 argued, dynamic, innovation-driven growth requires a nimble government that lowers transaction costs, facilitates coordination—by, for example, helping to set technological standards—and provides public goods such as infrastructure. Under the CDP, rather than simply "stepping back," the federal government did just that.[12]

---

[11] The deeper historical causes of this paradigm shift are complex and generally beyond the scope of this book. However, I do explicitly investigate the origins of the antitrust revolution—a key pillar of the CDP—in Chapter 7. There, I challenge the conventional wisdom that the shift to the "Rule of Reason" was merely the result of ideological capture by the Chicago School (see, e.g., Khan 2017). Instead, I argue it was largely an organic, pragmatic evolution by the judiciary, compelled by the need to coherently adjudicate the increasingly complex business practices of the modern economy.

[12] After reading this paragraph, some readers may balk at my depressing—forgive the pun—depiction of the statist-populist era and argue that the New Deal and its offspring policies were also in the business of solving market failures. They would indeed be partially correct in voicing this view. Many New Deal measures—such as the creation of the Federal Reserve (1913) and infrastructure projects like the Tennessee Valley Authority (TVA)—were indeed framed as

The Long and Winding Road to the CDP

The path to embedding the CDP in US regulatory practice was long and circuitous, shaped by a sequence of statutes, executive directives, and landmark court decisions that gradually reinforced evidence-based decisions in the service of innovation.[13] Under this aegis, policymaking was increasingly aimed at improving dynamic efficiency—or at minimum, insulating the innovation ecosystem from politicized efforts to freeze economic change in place. It all started with key institutional reforms.

In *Humphrey's Executor* (1935), the Supreme Court recognized the value of insulating certain agencies from fleeting political pressures. In that case, the Court upheld a provision of the Federal Trade Commission Act that barred the president from removing FTC commissioners except "for inefficiency, neglect of duty, or malfeasance in office"—a protection that, the justices held, was consistent with Congress's view of the FTC as a "quasi-legislative" and

---

technocratic solutions to systemic failures; and, and the TVA's provision of public infrastructure successfully "crowded in" private manufacturing by lowering input costs (see Kline and Moretti 2014), solving a genuine coordination failure. However, a rich literature in public choice economics argues that even the most technocratic-seeming reforms of the era—including the Federal Reserve, the Glass-Steagall Act, and the Social Security Act—were often driven by interest-group bargaining and resulted in significant rent-creation (e.g., Friedman and Schwartz 1963; Benston 1990). And regardless of their ultimate effects, they were not rooted in cost-benefit analyses that commenced from first principles (see Sunstein 2002).

[13] To tell this story (see below) I drew closely on Breyer et al. (2022), the canonical source for the history of the administrative laws and court decisions that strengthened them.

"quasi-judicial" body rather than an arm of the executive. By shielding the FTC from summary removal, Humphrey's Executor ensured that career experts could conduct long-range studies, collect data, and engage in fact-finding free from the threat of abrupt ouster.[14]

While Humphrey's was what legally allowed Congress to vest "independent" executive agencies with expert, apolitical authority, in 1946 Congress gave this court ruling teeth by enacting the APA, the Administrative Procedure Act (US Congress 1946). It imposed uniform procedures on federal rulemaking: agencies had to publish proposed rules, solicit public comment, and support final regulations with a "reasoned explanation" grounded in a robust administrative record. Moreover, courts must set aside agency actions found to be "arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law" (ibid: 243). To withstand judicial review, agencies must examine relevant evidence, connect facts to conclusions, and provide satisfactory explanations for their actions. The APA also requires agencies to build an administrative record including all data, studies, and comments used in formulating rules (ibid). Taken together, these features of the APA laid the procedural and evidentiary foundations for systematic cost–benefit analysis.

Congress enacted the APA in response to the expansion of federal agency authority during the New Deal era (see Shepherd 1996). The law was aimed at establishing consistent administrative processes, promoting transparency, and protecting individual rights while

[14] Later rulings—most notably *Free Enterprise Fund* (2010) and *Seila Law* (2020)—narrowed Humphrey's reach by striking down double-layer removal protections for certain financial regulators; moreover, President Trump challenged the independence of executive agencies and the constitutionality of Humphrey's Executor at the start of his second term.

balancing agency expertise with public oversight. While this reaction was squarely within the populist-statist era as outlined above, it represented the seeds of that paradigm's eventual demise. The APA embedded safeguards that came to underpin the CDP innovation-friendly regulatory order.

In codifying the APA's notice-and-comment regime and its "arbitrary and capricious" standard, Congress effectively institutionalized solutions to the collective-action problems I outlined in Chapter 2. By requiring agencies to publish proposed rules and solicit public input, the APA created a natural focal point—standardized procedures and fixed comment windows—that resolved pure coordination dilemmas by making clear, shared signals about when and how rules would be set.[15] By mandating a reasoned explanation of every significant policy choice, it imposed a fair-split dynamic: agencies had to lay out competing alternatives and justify their chosen path, thereby realigning distributional incentives and preventing any single stakeholder from capturing the process without public accountability.

Finally, by furnishing a robust administrative record that courts could review against an "arbitrary and capricious" test, the APA transformed one-shot commitment failures into iterated exchanges: once a rule survived judicial scrutiny based on logic and evidence, regulated parties could reliably anticipate its stability, turning what would otherwise be a Prisoner's-Dilemma scramble into a sustainable, mutually enforceable equilibrium.

---

[15] Indeed, the E-Government Act of 2002 created Regulations.gov to streamline notice-and-comment rulemaking by establishing a centralized electronic docket for proposed rules and public comments—this made it significantly easier for geographically dispersed stakeholders to coordinate on a single focal point (see US Congress 2002).

Fast forward some and, beginning in the early 1990s, the CDP infrastructure experienced a Cambrian explosion. A flurry of executive orders (EOs) and court decisions blessed regulatory cost-benefit analysis; electronic rulemaking expanded public participation, increasing the volume and diversity of evidence agencies must consider; and courts began to scrutinize personnel changes and policy shifts to ensure they aligned with reasoned explanations rather than arbitrary factors. I now turn to evaluating how these overlapping developments enhanced the APA's evidence-based framework.

Late 20th century Inflection Point

Several judicial decisions in the late 1970s and 1980s presaged 1990s executive actions that were instrumental to consolidating the CDP.

In Vermont Yankee Nuclear Power (1978), the Supreme Court held that courts cannot impose ad-hoc procedures beyond those specified in the APA, thereby channeling challenges back to the APA's built-in notice-and-comment and "arbitrary and capricious" standards. This meant that the judiciary was barred from inventing burdensome new requirements to stall agency action, effectively clearing the procedural path for regulators to implement complex, technical policies without fear of unpredictable judicial interference.

In Motor Vehicle Manufacturers Association (1983), the Supreme Court struck down an attempt by regulators to revoke a requirement for airbags. The Court held that the agency had failed to consider the data or offer a logical justification for the reversal, effectively establishing that regulators could not change rules on a political whim; they had to justify their decisions with a "reasoned explanation" grounded in hard evidence.

In Chevron (1984), the Court handed a decisive victory to technocratic expertise. The justices ruled that whenever a statute is ambiguous, courts must defer to the agency's reasonable

interpretation rather than substituting their own views. This decision effectively empowered agencies to use technical data and cost-benefit analysis to fill in the gaps of broad laws, safe in the knowledge that generalist judges would not second-guess their scientific conclusions (see Sunstein 2013).

By the early 1990s, the White House formalized this evidence-based revolution. President Clinton's EO 12866 mandated that any executive agency rule deemed "economically significant"—defined as having an annual impact of $100 million or more—must undergo centralized review by the Office of Information and Regulatory Affairs (OIRA). To pass muster, agencies were required to follow a strict five-step protocol: (1) define the specific problem to be solved; (2) identify all reasonable alternatives, including a "no action" baseline; (3) quantify costs and benefits to the fullest extent feasible; (4) demonstrate that the chosen regulation yields net benefits; and (5) transparently address public input (Clinton 1993). This order effectively transformed OIRA into the federal government's gatekeeper, ensuring that no major intervention proceeded without a clear economic justification.

By channeling significant executive rulemakings through OIRA, EO 12866 created a centralized mechanism for ensuring consistency, preventing poorly substantiated regulations, and giving stakeholders a clear rationale for why a rule was needed (see Sunstein 2013). Compared to the earlier Reagan-era framework (EO 12291), which drew thousands of draft rules per year into centralized review (see Reagan 1981a), EO 12866 focused OIRA's attention primarily on "significant" regulations (see Clinton 1993)—bringing the annual review flow down to the order of hundreds rather than thousands, with the "economically significant" subset smaller still (Copeland 2013). This emphasis on rigorous analysis fed directly into the National Economic Council (NEC), which served as the White House's strategic "war room," marshaling agency

experts to produce unified, data-driven policy recommendations, while the Council of Economic Advisors (CEA) acted as an external reality check, bringing the latest academic research into policy deliberations (see Destler 1996).

A decade later, the evidence-based revolution became fully consolidated. In 2003, the Office of Management and Budget (OMB) introduced Circular A-4 (OMB 2003). Acting as the technical "manual" for EO 12866, it provided agencies with standardized, rigorous methodologies for conducting Regulatory Impact Analyses (RIA). It stripped away ambiguity, specifying exactly how regulators must calculate discount rates, assess risk, and value future benefits, thereby ensuring that cost-benefit analysis became a disciplined science rather than a bureaucratic exercise.

Indeed, fundamentally departing from how RIA had been conducted hitherto, Circular A-4 was very specific about how regulators should engage in this exercise.[16] First, they should define the regulatory baseline by describing what would happen without the new rule. Second, regulators should choose appropriate time horizons and discount rates by specifying the

[16] Prior to Circular A-4, agency analysis was governed by the OMB's 1996 "Best Practices" guide, which offered general principles rather than binding technical standards (see OMB 1996). Consequently, the pre-2003 era was often characterized by methodological inconsistency: different agencies (or even the same agency across different rules) utilized varying discount rates, time horizons, and valuations of statistical life to justify preferred policy outcomes. By mandating uniform technical standards, Circular A-4 eliminated this "shopping for models" and forced agencies to subject their regulations to a unified, government-wide economic yardstick (see Dudley 2020).

evaluation period and applying both 3 and 7 percent real discount rates. Third, they should categorize and monetize policy effects by separating direct compliance costs from indirect effects and assigning monetary values where possible. Fourth, regulators should conduct sensitivity and uncertainty analysis through "what-if" scenarios to understand how robust conclusions are to variation in critical parameters. Finally, they should compare mutually exclusive and non-mutually exclusive options through side-by-side comparisons or appropriate aggregation of complementary measures (see ibid).

To put flesh on how Circular A-4 transformed the RIA process, consider the EPA's 2014 "Tier 3" tailpipe emissions standards for passenger cars, which were deemed "economically significant" under EO 12866 (see EPA 2014). The EPA first established a baseline by projecting future vehicle emissions and associated health impacts under existing Tier 2 vehicle and fuel standards, then evaluated tightened Tier 3 vehicle standards and a 10 parts per million (ppm) average gasoline sulfur standard against the status quo. For the cost-benefit analysis, EPA modeled changes in ambient ozone and other pollutants, estimated avoided premature mortality and morbidity outcomes, monetized these health gains using 3 percent and 7 percent discount rates, and compared them to projected compliance costs for automakers and refiners, including per-vehicle cost increases and per-gallon fuel costs.

Then, using these discount rates, the agency compared discounted benefits and costs across options and found that a 10 ppm average sulfur standard, combined with Tier 3 vehicle standards, delivered the needed emission reductions and the largest positive net benefits among the sulfur options it considered. EPA developed "high" and "low" benefit scenarios by varying key epidemiological and valuation assumptions and conducted cost sensitivities, concluding that the Tier 3 program remained strongly net-beneficial under a wide range of parameter choices.

The proposal and supporting analysis were issued for public comment, underwent interagency review, and were revised before EPA finalized the rule and the accompanying RIA.

Finally, several statutes, judicial decisions, and EOs helped consolidate this rigorous RIA process.[17] In Mead Corp. (2001), the Court clarified that, generally, only agency interpretations delivered through notice-and-comment rulemaking or formal adjudications should enjoy Chevron deference—pressuring agencies to ground policy choices in reasoned, record-based proceedings.[18] The E-Government Act of 2002 mandated that agencies post all proposed rules and their supporting materials online and provide an electronic docket for public comments (see US Congress 2002). Several executive branch agencies, working with the OMB, launched the first unified federal rulemaking portal called Regulations.gov in 2003, allowing the public to view rule proposals and submit comments electronically (GAO 2003). And EO 13563, issued in 2011, built on EO 12866's seminal cost-benefit framework by directing agencies to "promote

[17] For these developments see Warren (2020).

[18] This threshold inquiry is often called "Chevron Step Zero." If an agency action fails this test—for example, if it is issued through a mere policy letter rather than a formal process—it does not receive binding Chevron deference (see Chevron 1984). Instead, it receives only "Skidmore deference," meaning the court will follow the agency's view only to the extent that it has the power to persuade based on the thoroughness of its logic and expertise (see Skidmore et al. 1944). This doctrinal shift created a strong legal incentive for agencies to utilize rigorous notice-and-comment procedures to ensure their regulations would withstand judicial scrutiny (see Merrill 2002).

and facilitate public participation" and to use the best available techniques to quantify both benefits and costs, emphasizing the retrospective review of existing rules (Obama 2011).

The Strong Bipartisan Consensus Behind the CDP

Enjoying widespread consensus, during the CDP's heyday it transcended ideology and partisanship. In the runup to the Fourth Industrial Revolution, policymakers from both sides of the aisle believed that evidence-based policy in pursuit of innovation would make America more prosperous, stronger, and more influential (Block and Keller 2011; O'Mara 2019). Under that aegis, they supported not only market-oriented reforms and deregulation, but also proactive government involvement in fostering innovation and addressing market failures (see Mazzucato 2013; Block 2008). The proof is in the pudding: Between 1981 and 2019, OIRA reviewed over 2,500 "economically significant" regulations—each requiring an assessment of costs and benefits (quantified where feasible)—and the volume of such evidence-based rules persisted across presidencies held by both parties (OIRA 2024; Dudley 2020).

Importantly, different presidents built on their predecessors' commitment to the CDP and helped consolidate its nuts and bolts. While EO 12866 was issued by President Clinton (D) in 1993, it drew on OMB traditions stretching back to the Reagan (R) and Bush 41 (R) administrations (see Katzen 2011); later, Bush 43 (R), with EO 13422 (Bush 2007), and Obama (D), with EO 13563 (Obama 2011), each reaffirmed the centrality of cost–benefit analysis and public engagement in rulemaking. Even Trump's EO 13771 (Trump 2017), which imposed a

"two-for-one" requirement mandating that agencies repeal two existing regulations for every new significant one issued, operated through OIRA's CBA process.[19]

Moreover, the judicial confirmations that put judges on federal courts during this era ratified and strengthened the CDP. Across multiple Supreme Court and D.C. Circuit nominations, both Republican and Democratic senators routinely affirmed that judges should respect Chevron deference and uphold the APA's "arbitrary and capricious" standard —that is, that they should defer to agency expertise on technical matters rather than substitute their own judgment—another signal that evidence-based rulemaking commanded broad support (see Gaffney 2024; Klobuchar 2024).[20]

---

[19] Beginning with the first Trump administration, the CDP experienced significant challenges that included the notion that (i) executive agencies should retain independence from the president and (ii) regulations should be disciplined by cost benefit analysis.

[20] Beginning in 2022, however, the Supreme Court began to rein in Chevron's deference and tighten judicial oversight of agency rulemaking. In West Virginia et al. (2022), the justices invoked the "major questions" doctrine to hold that regulations of vast economic or political significance demand clear congressional authorization, thereby forcing agencies to tether their CBAs more explicitly to statutory text. Two years later, in Loper Bright (2024), the Court overruled the Chevron (1984) ruling itself, declaring that when a statute is ambiguous courts must interpret it directly rather than rubber-stamp an agency's "reasonable" reading. These decisions expose agencies' cost–benefit studies to independent judicial scrutiny, raising the evidentiary bar for every high-stakes rule.

The upshot is that, throughout the CDP era, both Democrats and Republicans repeatedly turned to the same toolbox—notice-and-comment procedures, centralized OIRA review for executive agencies, and RIAs—to discipline the regulatory process through cost-benefit analysis. That is, until several cracks began to show during the first Trump administration (see Chapter 12).

Consensus about the Ends, not Just Means

Beyond their shared consensus about the regulatory process's nuts and bolts, both political parties also supported a host of policies that promoted innovation and directly fed into the Fourth Industrial Revolution. Consider some of the examples with strong bipartisan support that I will flesh out further ahead, both in this chapter and Chapter 8. First, the Bayh–Dole Act, which was co-authored by Senator Birch Bayh (D-IN) and Senator Bob Dole (R-KS), transformed how federally sponsored inventions were commercialized, passed the Senate unanimously (US Congress 1980b). Second, the Stevenson–Wydler Technology Innovation Act, championed by Senator Adlai Stevenson III (D-IL) and Representative John Wydler (R-NY), codified technology-transfer offices at national labs, also secured unanimous support (see US Congress 1980a). Third, Section 230 of the Communications Decency Act, crafted and shepherded through Congress by Representative Chris Cox (R-CA) and then-Representative Ron Wyden (D-OR), passed with near-unanimous support in the House (US Congress 1996). Finally, the 21st Century Cures Act, which unlocked $6 billion for biomedical innovation, passed with overwhelming votes in both legislative chambers (see US Congress 2016).

What's more impressive is that the CDP succeeded during an era of increasing political polarization. As McCarty et al. (2016) argue, beginning in the 1970s political polarization increasingly hollowed out the US's moderate center and pushed both parties toward ideological

extremes. More concretely, Harbridge-Yong (2015) reports that among final-passage votes on House bills, the bipartisan share—where majorities of both parties agree—declined from about 80% in 1973 to about 40% by 1994. Moreover, this partisan divide was not merely cosmetic; it reflected a deep ideological sorting that eventually reached levels of polarization not seen in the US since the post-Civil War Reconstruction era (McCarty et al. 2016).

Public sentiment mirrored the cross-aisle consensus about the government's role in promoting innovation. National Science Board (NSB) data shows consistent majority support for federal funding of basic scientific research across several decades. In 2022, for instance, 88% of US adults agreed that scientific research deserves federal government support (NSB 2024). Similarly, as late as 2025, a nationwide poll conducted by the Association of Science and Technology Centers (ASTC) found that about nine-in-ten respondents viewed federal investment in STEM as vital for future economic success, with large majorities admitting they directly benefit from federally supported science (ASTC 2025).

THE GOVERNMENT SETS UP A COMMERCIALIZATION FRAMEWORK

Overwhelming bipartisan consensus about the benefits of the government's role in fostering innovation allowed policymakers to move beyond high-level theory and construct the specific legal and regulatory scaffolding necessary to sustain the digital economy.

Most basically, Washington, D.C. nurtured innovation through a supportive fiscal environment that made high-risk innovation financially viable. A combination of reductions in capital gains taxes (which increased the supply of venture capital) and the 1981 Research and Experimentation Tax Credit (which subsidized invention) fundamentally altered the calculus for private firms (see O'Mara 2019; Guenther 2022). Importantly, the federal government also updated the tax code to reflect the digital age: accelerated depreciation schedules allowed

telecommunications firms and data centers to rapidly write off the costs of fiber-optic networks, servers, and routers (Gravelle 2014). This effectively lowered the price of building the physical internet, incentivizing the private sector to deploy the massive infrastructure necessary for the Fourth Industrial Revolution.

In Chapter 7, I examine the role of antitrust in fostering innovation. I detail how the judiciary, responding to the practical complexities of modern markets, organically shifted from rigid "per se" prohibitions on a host of business practices and merger attempts to the flexible "Rule of Reason."[21] This evidence-based framework prioritized consumer welfare and innovation over competitor protection, effectively legalizing the vertical integration and massive scale that allowed digital platforms to build the data ecosystems necessary for the AI revolution.

In Chapter 8, I will explore how Section 230 of the Communications Decency Act provided online platforms with immunity from liability for user-generated content, enabling them to host vast amounts of information without fear of legal repercussions and ultimately stimulating the growth of vibrant digital communities and the rise of social media that employed algorithmic techniques to bolster user engagement. This was also critical to fostering demand-side economies of scale in digital markets.

In that chapter I also briefly examine how this legal framework was paired with the physical infrastructure reforms of the 1996 Telecommunications Act—such as market-based

[21] As I will outline in that chapter, such an assessment first asks if the arrangement or behavior in question engenders some harm to competition, such as keeping rivals out of a market and, if so, if it is reasonably necessary to achieve pro-competitive benefits that outweigh those harms; for example, by facilitating large-scale R&D collaborations that spur innovation.

spectrum auctions and broadband deregulation. By providing a reliable framework in the 1990s through which high-speed 4G and 5G pipes would later connect mobile users during the 21st century, the US federal government made the perpetual streaming of geocoded information a reality.

In the rest of this chapter, I investigate how under the aegis of the CDP policymakers made dramatic reforms to IPRs and liberalized international trade. I will also examine other reforms that made digital commerce easier and less risky.

As I argue later in the book, the convergence of all these developments fostered the massive, real-time datasets required to train the foundational models of the AI age. By ensuring that the "new oil" of data would be legally protected, as well as produced, extracted, and processed at a global scale, the CDP architects inadvertently seeded the Fourth Industrial Revolution.

IPRs Take Center Stage

IPRs are a red thread that splices across this book and its supplementary online appendix: in Menaldo (2026, Section 1) I tender several examples of how creative IP licensing strategies around semiconductors accelerated technological diffusion and helped consolidate distinct industrial revolutions. In Chapter 5, I'll deploy a quantitative network analysis to map the technological lineages of the Third Industrial Revolution. There, I also provide a detailed technical primer on US patents during the 1976-2006 analysis period—including specific standards governing novelty, non-obviousness, and utility. In this chapter, I focus on the legal and regulatory changes that made IP rules stronger and more predictable during that era.

Before getting into the nitty gritty of the key reforms, let's review some basics. The American patent system operates within a well-established legal framework, with IPRs enshrined

in Article 1, Section 8 of the Constitution. Congress creates the foundational laws governing patents (see, e.g., US Congress 1952), while the USPTO develops these laws' implementing regulations (NARA 2024). To guide patent examiners in interpreting and applying these laws consistently, in 1949 the USPTO created the Manual of Patent Examining Procedure (MPEP), which it has revised several times since (see USPTO 2024).

Watershed IP Reform Era

Beginning in the late 1970s and extending through the 1990s, policymakers significantly strengthened American IPRs. Table 3.2 outlines five pivotal reforms to the pre-CDP IP system: the creation of the Court of Appeals for the Federal Circuit (CAFC) to unify patent appeals; the gradual legal recognition of software patents; the Bayh-Dole Act's incentivization of university technology transfer; the globalization of IP standards via TRIPS and the WTO; and federal support for organizations managing Standard Essential Patents (SEPs). Together, these changes were a welcome shot in the arm to the inventors, entrepreneurs, manufacturers, and distributors involved in commercializing new technologies that underpinned the digital economy I discuss in ensuing chapters. Before I take up each of the items in Table 3.2 separately, I briefly recount the seminal court decisions, laws, and executive branch actions that led to this IPR revolution.

Table 3.2 Key CDP Changes to IP

1. Creation of the CAFC (1982)	
What It Did	Unified patent appeals by consolidating them in one specialized court; replaced a fragmented system where different circuits had inconsistent rulings.

Why It Mattered	Ensured uniformity and predictability in patent rulings, reducing legal uncertainties for innovators; encouraged higher R&D investment, especially in tech and biotech. Bolstered the 'creative' side of creative destruction by giving inventors a stable legal environment; fostered competitiveness by pushing incumbents to invest in innovation.
2. Patents for Software	
What It Did	Software patentability was historically ambiguous; from the 1980s-1990s, CAFC rulings (e.g., Alappat 1994) and USPTO guidelines gradually recognized software as patentable.
Why It Mattered	Software patents became key to tech startups and established firms, supporting investment in code-based R&D; fueled the rise of the PC era and early web-based businesses. Expanded the scope of intangible IP, spawning robust digital entrepreneurship; enabled startups to disrupt incumbents by protecting innovative software solutions.
3. Bayh-Dole Act (1980)	
What It Did	Allowed universities, nonprofits, and small businesses to own patents on federally funded research; before, the federal government owned such patents, which often went uncommercialized.
Why It Mattered	Encouraged direct commercialization of academic discoveries (e.g., biotech, engineering); led to tech-transfer offices, spin-offs, and

	public-private R&D collaboration. Transformed universities into innovation engines, accelerating the 'creative' aspects of technological change; fostered an entrepreneurial culture in research institutions.
4. TRIPS Integration into the WTO (1995)	
What It Did	TRIPS (Trade-Related Aspects of Intellectual Property Rights) set minimum IP standards for WTO members, covering patents, copyrights, trademarks, etc.
Why It Mattered	Internationalized U.S.-style IP rules, helping American firms protect innovations abroad; reduced IP theft risk, promoting cross-border licensing and investment. Globalized the intangible economy by guaranteeing IP protection in export markets; enabled offshoring of lower-value tasks while retaining high-value R&D in the U.S.
5. Federal Support for Organizations Managing SEPs	
What It Did	Government agencies (e.g., NIST) legitimized voluntary standard-setting bodies and supported FRAND licensing regimes; SEPs are critical to tech standards (e.g., Wi-Fi, 4G LTE).
Why It Mattered	Ensured interoperability across firms while protecting core IP; facilitated coordination and faster adoption of shared technologies. Supported platform-based ecosystems (e.g., smartphones, IoT), enabling firms to build on shared standards; reinforced an intangible-driven market while incentivizing innovation.

Notes: CAFC = Court of Appeals for the Federal Circuit; USPTO = United States Patent and Trademark Office; WTO = World Trade Organization; TRIPS = Trade-Related Aspects of Intellectual Property Rights; SEPs = Standard Essential Patents; FRAND = Fair, Reasonable, and Non-Discriminatory licensing terms; IoT = Internet of Things; CDP = Creative Destruction Paradigm. NIST = National Institute for Standards and Technology

In *Studiengesellschaft Kohle* (1981), the D.C. Circuit reversed a lower court's order that would have mandated compulsory licensing. By holding that a patentee's restrictive licensing terms were generally within the scope of the patent grant, the decision signaled a retreat from the aforementioned "Nine No-Nos" enforcement policy that had previously treated such restrictions with deep suspicion (see Gilbert et al. 1997).

In 1984, the Semiconductor Chip Protection Act created *sui generis* protection for the layouts of integrated circuits known as "mask works" (US Congress 1984a; Samuelson and Scotchmer 2002). This addressed a critical gap where neither patent nor copyright law adequately prevented competitors from copying chip designs.[22]

By the mid-1990s, antitrust authorities had formally abandoned the presumption that patents automatically confer their holders with market power. The DoJ and FTC's 1995 Antitrust Guidelines for the Licensing of Intellectual Property (building on principles articulated in the

[22] Moreover, key patent lawsuits helped cement stronger IPRs by punishing infringers. Between 1985 and 1986, Texas Instruments successfully sued Japanese and Korean firms for infringing its integrated-circuit patents; in 1986, Polaroid's landmark victory over Eastman Kodak yielded a one-billion-dollar award (see Polaroid 1986).

DoJ's 1988 international guidelines) explicitly declared that IP should be treated like any other property and that well-structured licensing arrangements are generally pro-competitive (see DoJ 1988; DoJ and FTC 1995).

During the 21st century, this pro-patent regime continued. In the late 1990s and early 2000s, the DoJ issued "Business Review Letters" that effectively endorsed patent pools for standards like MPEG-2—which aggregated essential video-compression patents to simplify licensing for device manufacturers—provided they included safeguards to preserve competition (see DoJ 1997). The Supreme Court's decision in Illinois Tool Works (2006) officially abjured the notion that the mere existence of patent rights was tantamount to market power in antitrust cases involving "tying"—as I explain in Chapter 7, in this scenario a seller conditions the sale of one product (the tying product) on the buyer also purchasing a second product (the tied product).

Finally, the Federal Circuit's CSIRO (2015) and Exmark (2018) rulings refined the calculation of damages in IP infringement lawsuits. In the first case, the court rejected a rigid requirement to always start with the smallest salable patent-practicing unit—namely, the smallest distinct component (like a chip) that embodies the invention—allowing judges to instead consider real-world licensing negotiations as evidence. In the second one, the court ruled patentholders may utilize the value of the entire product as a royalty base (which means that the royalty calculation starts with the sales price of the finished good), provided that the royalty rate is scaled down to reflect only the value of the patented improvement.[23]

[23] To understand the logic here, consider the facts of the case itself, which concerned a patented "baffle" (a metal shield) used on a commercial lawnmower deck. Under the old, rigid "Smallest Salable Patent-Practicing Unit" approach, the damages calculation would have been forced to

This method is preferred by those who are concerned about innovation incentives because it recognizes that a single, small component (like a microchip) can drive the market demand for a complex, expensive system. It thus allows licensors to be compensated for their specific contribution to a product's commercial success, albeit without being overpaid for the unpatented components they did not invent (see, e.g., Chien et al. 2019).

Further below, when I discuss the dispute between Apple and Qualcomm over how to divvy up the profits from the sale of iPhones when the latter's modem chips are an essential component of that device, I will revisit the importance of legally recognizing this basic economic fact: that in modern supply chains, a single component often determines an outsized share of a finished product's market value.

Court of Appeals for the Federal Circuit

The creation of the Court of Appeals for the Federal Circuit (CAFC) in 1982 was the IPR revolution's most powerful salvo. Before its inception, patent appeals were heard in the various regional circuit courts, resulting in a fragmented legal landscape where inconsistent rulings

---

start with the price of the lawnmower deck (e.g., $1,000) rather than the whole mower (e.g., $10,000), on the theory that the patent didn't cover the engine or wheels (see Exmark 2018; Clayton and Stone 2018). The Exmark ruling allows the patent holder to use the full $10,000 price as the base—acknowledging that the baffle might be the main reason a customer buys that specific mower—but requires apportionment to ensure fairness. In practice, this is a mathematical see-saw: if the court allows the use of the larger Royalty Base ($10,000), it generally insists on a lower Royalty Rate (percentage) to mathematically strip out the value of the non-infringing parts (engine, wheels).

stifled innovation. Depending on geographic jurisdiction, similar patent disputes often yielded opposite outcomes, leading to forum shopping—plaintiffs strategically filing lawsuits in circuits known to be favorable to their specific claims (Atkinson et al. 2008).

Recognizing that non-uniform federal appellate decisions created systemic uncertainty, the 1975 Hruska Commission recommended a National Court of Appeals to resolve inter-circuit conflicts. While the Commission did not explicitly call for a patent-only court, it warned against the dangers of legal Balkanization (CRFCAS 1975). Informed by the Commission's findings, lawmakers contemplated providing a stable, uniform body of law that would give investors the confidence to fund risky, capital-intensive technological projects without fearing that a hostile or inexperienced court might invalidate their IP (US Congress 1981). Their reform efforts culminated in the Federal Courts Improvement Act, which established the CAFC by merging the US Court of Customs and Patent Appeals with the Court of Claims (US Congress 1982).[24]

The impact was immediate. Legal centralization reshaped the patent system in three ways.

First, it eliminated the regional disparities in patent validity. By unifying precedent, the CAFC ended the era where a patent might be valid in Texas but invalid in New York. This harmonization significantly increased the commercial value of patents by providing a predictable, nationwide IP enforcement standard (Dreyfuss 1989).

[24] The CAFC has nationwide jurisdiction to hear appeals in specialized cases, including patent law. It is one of 13 appellate courts that sit below the US Supreme Court. Unlike trial courts, appeals courts consist of three judges and do not use a jury system when reviewing cases.

Second, the CAFC raised the bar for invalidating patents by eliminating "hindsight bias"—the tendency to look at a completed invention and assume it was obvious all along. In cases such as Deuel (1995) and Rouffet (1998), the court ruled that a patent examiner could not reject a patent simply because its separate components already existed.[25] Instead, they imposed the strict "Teaching-Suggestion-Motivation" test: to reject an invention, the examiner had to find specific documentary evidence from the past that explicitly suggested combining those exact parts. This created a high evidentiary hurdle for challengers (see Durie and Lemley 2008). And although the Supreme Court eventually relaxed this test (see KSR 2007), the CAFC's strict application of the doctrine during this critical period in technological development (see Chapters 4-8) effectively firewalled patents that were already granted from being second-guessed, giving investors the certainty they needed to fund complex new technologies.

Third, it strengthened IP infringement remedies. Prior to the CAFC, courts granted patentholders preliminary injunctions—a court directing law enforcement to block an accused patent infringer's products from entering the market or seize them at the border—under an "unusually stringent" approach that sometimes required that patents be "beyond question" valid and infringed (Shapiro 1993). What this meant was that preliminary injunctions were exceedingly rare, effectively allowing accused infringers to continue selling competing products throughout the duration of a trial that could last for years. In the post-1982 era, the CAFC

[25] These cases were disputes between patent applicants and the USPTO, rather than lawsuits between two private parties. In these cases, the court was effectively overruling examiners who had rejected patent applications.

applied a more standard "likelihood of success" framework, making injunctive relief a more viable tool for patentholders seeking to protect their inventions from infringement (see ibid).[26]

Together, these reforms vouchsafed inventors and investors with stronger IPRs that were more certain and more readily enforced. This, in turn, coincided with a surge in activity: total patent applications at the USPTO grew from roughly 112,000 in 1980 to over 315,000 by 2000—a nearly threefold increase (USPTO 2021). As the system digested this surge, it set the stage for the massive expansion of IP rights into the key technologies of the digital revolution, software and semiconductors. This is a topic I will explore in greater detail below, starting with software.

Legal Recognition of Software Patents

Long before software patents became widespread, software companies such as Microsoft protected their early products through copyright, licensing, trademarks, and trade-secret law.

In the 1940s and early 1950s, software patentability was not a recognized concept. The USPTO treated computer “programs” as unpatentable methods or abstract ideas. Indeed, when the first commercial computers came to market shortly after World War II, there were no pure

[26] Over time, the CAFC refined its approach, emphasizing that injunctions are an extraordinary remedy rather than an automatic right. This shift was cemented by the Supreme Court in eBay (2006), which replaced the previous practice of granting nearly automatic permanent injunctions upon a finding of infringement with a mandatory four-factor equitable test. Under this "eBay standard," a plaintiff must demonstrate: (1) it has suffered an irreparable injury; (2) monetary damages are inadequate to compensate for that injury; (3) the balance of hardships between the plaintiff and defendant warrants an equitable remedy; and (4) the public interest would not be disserved by a permanent injunction (see eBay 2006; Fues 2007).

software patents in their IP portfolios. Instead, the most significant patents covered physical components like memory systems and architectural concepts. Instead, their most significant patents covered physical components like the machine's memory system and its architectural concepts.

In the next decade, courts continued to treat algorithms as abstract mathematical ideas outside the scope of patents. While a few daring firms quietly pressed the USPTO to protect their computer code beginning in the mid-1960s, they were rebuffed. For example, Applied Data Research engineer Martin Goetz recalled that his 1968 submission for a sorting-algorithm patent was rejected by the USPTO until he reframed it as part of a "digital data-processing system" (Bessen and Meurer 2008: 73). And when the Supreme Court confronted pure software claims in *Gottschalk* (1972)—an algorithm for converting binary-coded decimal—it held that unembodied programs were not statutory subject matter, a decision reinforced in *Parker* (1978).

The 1970s were no better. From the moment Altair BASIC's source code—the seminal interpreter for the first PC—was written to tape in 1975, it was safeguarded by copyright as a "fixed" work (see US Congress 1976b: Sect. 10). Rather than patenting the BASIC language, Microsoft agreed to an exclusive license with Micro Instrumentation and Telemetry Systems to control distribution (Wallace and Erickson 1992). Microsoft also kept its interpreter internals as trade secrets by withholding source code and integrating proprietary performance routines (Wallace and Erickson 1992).[27] But this was a costly distraction that drew the company's

[27] Interpreter internals refer to the underlying code and algorithms that enable a programming language interpreter, such as Altair BASIC, to process and execute instructions. Monte Davidoff's floating-point routines are specialized code for handling decimal calculations.

attention away from its comparative advantage developing software, instead of finding clever ways to protect it from IP infringers (Gates 1976).

Fortunately for Microsoft, by the 1980s a new IP paradigm around software began to emerge, albeit fitfully: the CAFC began to uphold certain software-related patents. In *Freeman* (1978), the CAFC's predecessor, the Court of Customs and Patent Appeals, held that even though the claimed process ran on a general-purpose computer, a system for typesetting alphanumeric information could be patented if it produced a "useful, concrete and tangible result" (see Quinn 2014). The Supreme Court's decision in *Diamond* (1981) then ostensibly cleared the way for software to be "hidden" under hardware metaphors. There, the Court upheld claims to a computer-implemented method of curing synthetic rubber because the steps recited the use of a digital computer in conjunction with physical-transformation steps, thereby treating the mathematical algorithm as patentable subject matter when applied in an industrial process.

In practice, however, during the 1980s and 1990s examiners and courts enforced the "magic words" doctrine. This meant that a brilliant piece of software would be rejected if claimed simply as "code," but accepted if the attorney added standard hardware phrases—such as "a general purpose computer configured to…" or "a computer-readable medium containing…"—to the claim. By demanding that software be linguistically tethered to a physical machine (even a generic one), this formalistic rule forced innovators to engage in a drafting game, delaying the day that the code itself was recognized as the true source of value (Cohen and Lemley 2001).

Throughout the early 1990s, the CAFC issued a series of decisions that incrementally expanded the patentability of software, which was reinforced by changes at the USPTO. In *AT&T* (1999), the court accepted recitations of computer "means for executing" a method that were more abstract, signaling that detailed hardware metaphors were no longer the sole path to

patentability (Crissinger 2000). Meanwhile, patent examiners grew more tolerant of claims for a "computer-readable medium"—a legal phrasing that allowed software to be patented as a physical object, like a floppy disk containing code. They also began accepting functional descriptions like "a processor configured to [perform task X]," meaning inventors no longer had to draft Byzantine diagrams enumerating every single integrated circuit element—the millions of microscopic transistors and logic gates inside the chip—to prove they had built a machine (see BitLaw 2025).

Finally, in Alappat (1994), the court faced a dispute over a seemingly simple visual improvement: a smoother line on a screen. The invention was designed for an oscilloscope—a standard laboratory instrument used by engineers to visualize electrical signals as waves. The problem was that early digital screens often rendered these curved waves with jagged, staircase-like edges—a visual glitch known as "aliasing." Alappat's software solved this using anti-aliased pixels, a technique that varied the brightness of the dots along the line's edge to trick the human eye into seeing a perfectly smooth curve.

While the USPTO initially rejected the patent, arguing this was merely a mathematical calculation (an abstract idea), the CAFC disagreed, ruling that the software effectively created a new, specific machine: a rasterizer—the hardware component that converts raw data into the visible grid of pixels on a screen. The court concluded that because the software physically transformed a general-purpose computer into a specialized tool capable of this specific visual task, it was a patentable invention rather than a disembodied concept (Alappat 1994; Kim 1995).

This case was seminal. It introduced a new doctrine that effectively overturned the USPTO's longstanding policy against software patents, paving the way for inventors to patent inventions that can be implemented in either hardware or software (Kim 1995). Indeed, by the

late 1990s, the CAFC made this the official litmus test for software and even business method patents (BitLaw 2025). As with all things property rights, greater certainty meant a more hospitable investment climate: Key players in the software industry noted that the *Alappat* decision offered them "an excellent opportunity to…protect the value of technological advancements to a far greater degree than previously permitted" (Abelson 1994: 18-19).

The USPTO moved quickly to align its internal procedures with this new judicial mandate. Following the 1994 public hearings on software patentability and a subsequent request for public comment in 1995 (see USPTO 1995), the agency concluded that the economic nature of software had fundamentally changed. Stakeholders presented data showing that coding was no longer a low-cost garage activity but a capital-intensive industrial process, with average development budgets surging past $10 million per project (see Hall and Ziedonis 2001). Confronted with evidence that the lack of property rights was distorting investment decisions, the USPTO updated its examination guidelines to reflect the new case law.

Then, the following year, it formally instructed patent examiners that they should evaluate software-implemented inventions just like any other process or machine, explicitly stating that a computer program could be patented if claimed as a series of steps, i.e., a process, or when tied to a machine-readable medium or programmed computer (see USPTO 1996). In short, while mere algorithms remained unpatentable, software that had a practical, functional application had become legally patentable.

By the late 1990s, software patents received full embrace. In 1998, the court dispensed entirely with the magic-words formality, holding that in that a financial algorithm for a hub-and-spoke cash management system was fully patentable (State Street Bank 1998). The implication was that any process producing a "useful, concrete, and tangible result" could be patentable,

including software-enabled methods" (ibid). More specifically, innovators in the software market could now secure patents for breakthroughs around graphical user interfaces, data compression algorithms, encryption protocols, network communication systems, and database management architectures.[28]

What Stronger Software Patents Wrought

Soon after, all manner of patents were granted for cryptographic algorithms, packet switching solutions for wireless communication, relational database query optimizers, indexing algorithms, and distributed storage systems (Allison and Tiller 2003). By the end of the 1990s, thousands of software patents were being granted each year, and many high-tech startups listed patent portfolios among their key assets when seeking funding (Bessen and Hunt 2007).

These developments had a huge impact during the twilight of the Third Industrial Revolution, when the smartphone economy blossomed, and eventually the Fourth Industrial Revolution, when AI ascended in importance. This was for three reasons.

First, stronger software patents aligned legal property rights with the new economic reality. As the value of technology shifted from physical hardware ("atoms") to intangible code ("bits"), the legal system had to evolve to protect where the actual value resided (see Lev 2001; Chapters 4-8 of this book). Without this shift, the core assets of the smartphone and AI

[28] To be sure, Alice Corp. (2014) later tightened scrutiny around "eligibility," shifting the court's focus from hardware metaphors to whether claims added an "inventive concept" beyond an abstract idea; the mere invocation of a "computer" or "network" would not rescue abstract software claims, despite the long-vanished "magic words" requirement (see Alice Corp.).

economies I explore in this book—from the signal-processing logic of 4G networks to the neural architecture of LLMs—would have remained legally defenseless.

Second, they catalyzed commercialization by reducing transaction costs—a core function of the patent system I detailed in Chapter 2. By delineating clear property boundaries around intangible code, patents transformed abstract algorithms into tradable assets. They not only provided collateral for venture capitalists to fund high-risk R&D (Mann and Sager 2007) but allowed firms to safely unbundle supply chains. With secure rights, specialized software startups could disclose their inventions to potential partners, licensees, or acquirers without fear of expropriation, thereby facilitating the complex business deals and specialization that defined the dot-com era (see Gans et al. 2002; Spulber 2015; Chapters 4-6 of this book).

Third, stronger software patents facilitated the complex licensing markets sustaining platform ecosystems. As I will argue below and in the ensuing chapters, interoperability made both the smartphone and AI revolutions possible, as both relied on thousands of different patented technologies working together. Clear rights over these digital building blocks sometimes allowed firms to cross-license, creating the stable legal architecture necessary for global digital platforms to scale (Arora et al. 2001; Graham and Vishnubhakat 2013).

Consider Microsoft again. When Windows 1.0 debuted in 1985, every executable, icon, and dialog layout carried a copyright notice, and the “Windows” mark was asserted as a trademark (Microsoft 1985). As the graphical user interface matured, however, Microsoft turned selectively to patents for novel system features. By the late 1990s, the Redmond-based company

had begun aggressively patenting user interface technologies and operating system functions, securing protection for features such as the "Taskbar" and menu structures.[29]

Although broad “look-and-feel” claims were rebuffed by the CAFC in Apple (1994), Microsoft and Apple eventually resolved their IP dispute—over the visual "look and feel" of their operating systems—by agreeing to a cross-license and settlement pact in 1997, a deal that involved both patents and copyrights. The upshot was concrete joint innovation, in that it ensured the sustained release of Microsoft Office for Macintosh, thereby guaranteeing critical file interoperability between the two rival platforms (see Markoff 1997).

The Bayh-Dole Act

Even spectacular inventions can stagnate when there is no repeatable path from lab to product. Bayh–Dole mattered because it tried to industrialize that path—making commercialization less dependent on heroic individuals or one-off institutional luck. By allowing universities and small businesses to claim ownership of inventions financially supported with federal research grants—while preserving the government’s “march-in” safeguards and enabling exclusive licensing when necessary—the Bayh–Dole Act aimed to stimulate a greater degree of technological commercialization (Mowery and Ziedonis 2001). Indeed, President Carter praised the legislation as promising “real benefits to the Nation's economic health by stimulating our people's innovative activity” (Carter 1980b).

---

[29] For example, take US Patent No. 5,825,357, "Method and system for providing a taskbar in a graphical user interface," which was filed by Microsoft Corporation in 1995 and issued in 1998 (see Malmud et al. 1998). This patent covers both the fundamental "Start" button and taskbar functionality that became a hallmark of the Windows interface.

In this vein, the Act fits into a broader, early-1980s push to make federally funded science easier to move into products. The Stevenson–Wydler Technology Innovation Act required large federal labs to create formal technology-transfer offices (ORTAs) and to treat transfer as part of their mission (see US Congress 1980b). Congress then expanded the toolkit with the Federal Technology Transfer Act of 1986, which authorized Cooperative Research and Development Agreements, a standardized contract form for lab–industry collaboration (US Congress 1986a)—this helped normalize public–private technology partnerships over the years that followed.

The Act's Impetus

Prior to 1980, inventions arising from federal government-funded R&D (for example, discoveries made in university laboratories with NIH or NSF grants) were typically owned by the government. Different agencies had inconsistent policies, but often the inventors and their institutions had to hand over patent rights to Washington, which would then offer nonexclusive licenses without any prior screening or forethought. In practice, this system led to thousands of government patents sitting on the shelf with no commercial development (GAO 1998).

During the late 1970s and early 1980s, economists and policymakers railed against the fact that federally funded research was "trapped" in university and government laboratories. For example, a US Senate Judiciary Committee report issued in 1980 (SCoJ 1980) documented dozens of promising compounds discovered through NIH grants that never reached clinical trials as private investors balked at underwriting the $100 million–plus needed for human testing (Cohen et al. 2002). Experts claimed that this "valley of death" not only delayed patients' access to new drugs but also deprived the American economy and society of the rightful fruits of innovation, as too many inventions subsidized by taxpayers failed to culminate in new products or firms (see Mowery and Sampat 2005). After a decade marked by stagflation and rising

industrial competition from Japan and Europe, a bipartisan group of highly influential US legislators warned that failure to translate public-sector research into marketable products contributed to the country's "economic malaise" (US Congress 1979a: 21).

Evidence-based Campaign & Bipartisan Support

Like the evolution of software patents, the Bayh–Dole Act of 1980 was the product of a deliberate, evidence-based campaign. Senators Birch Bayh and Bob Dole marshaled data to prove that a patenting "bottleneck" was stifling American ingenuity. The statistics were stark: the House Judiciary Committee reported that fewer than 5 percent of the roughly 28,000 patents in the federal inventory had ever been licensed (HJC 1980: 15; GAO 1998). Promising NIH-supported compounds often remained shelved because the lack of exclusive rights made the cost of clinical trials financially prohibitive. As the Senate Judiciary Committee observed, a non-exclusive license was tantamount to "no patent protection at all" (SCoJ 1980: 6).

This data provided the necessary "technocratic bridge" to unite the bipartisan coalition I previously identified in this chapter as the bedrock of the CDP. By framing the bill as a solution to industrial stagnation rather than a political spoil or an ideological pet project, the sponsors aligned conservative desires for decentralization with liberal goals for job creation and competitiveness (Stevens 2004). This consensus allowed the measure to bypass typical partisan friction, passing with overwhelming support during the 1980 lame-duck session (see US Congress 1980b).

The Act's Impact on Innovation

In unlocking a treasure trove of research that had been stashed and forgotten in American universities, the Bayh–Dole Act was dubbed as "[p]ossibly the most inspired piece of legislation to be enacted in America over the past half-century" (The Economist 2002: 3). By the late 1990s,

hundreds of universities had established active licensing programs, effectively transforming many research institutions into business incubators. While US universities received fewer than 400 patents annually prior to the Act, they subsequently experienced an exponential surge in activity: in the decade following its passage, the number of patents issued to institutions of higher learning more than doubled, and licensing revenues increased over 200-fold between 1980 and 1995 (Mowery et al. 2001; Sampat 2006).

Consider the Cohen-Boyer recombinant DNA technology, which became the gold standard for the technology transfer model solidified by Bayh–Dole (Feldman et al. 2007). Developed at Stanford University with federal funding, the invention was managed under an early institutional agreement that prefigured the Act. After its passage, Stanford granted dozens of non-exclusive licenses in the first year alone—each under standardized terms of roughly $10,000 upfront plus low single-digit royalties—to companies ranging from startups like Genentech to established firms (Feldman et al. 2007).[30] These licenses directly spurred industry-wide investment, helped launch over thirty new biotech firms, and spawned transformative products including synthetic insulin and new cancer therapies (Mowery et al. 2004).

Between 1996 and 2010, Roessner et al. (2012) estimate that university research licensed for commercial innovation contributed significantly to the U.S. economy. They calculated that, while the total economic impact on gross industry output was as much as $686.9 billion (in 2005 dollars), its effect on GDP was over $277 billion (in 2005 dollars).

International Efforts to Protect IP

[30] To appreciate the scale of this achievement, consider that in 1980, the entire US university system was granted only 390 utility patents total (NSB 2020: Table 5b-5).

While domestic political battles over trade never disappeared after World War II, the underpinning of US trade policy from the late 1970s on was largely technocratic and fully complemented the IP reforms outlined above. Whether it was the passage of NAFTA in 1994, or the formation of the WTO, or China's entry into this organization in 2001, the pattern was identical: the Executive Branch architected trade policy using data, expert analysis, and interagency consensus-building to pursue liberalization. It then persuaded Congress to codify these changes in law. And Congress obliged, usually stressing the need to protect American IP in the process.

In December 1994, President Clinton signed the Uruguay Round Agreements Act (URAA), implementing the agreements that transformed the General Agreement on Tariffs and Trade (GATT) into the WTO, which entered into force on January 1, 1995. Membership in the organization obligated nations to uphold the Trade-Related Aspects of Intellectual Property Rights (TRIPS). This substantially strengthened global patent protections by imposing uniform, stringent IP rules across developing economies (see Maskus 2000). The US also enhanced global IP protections through bilateral and multilateral agreements, such as the North American Free Trade Agreement (NAFTA).

As with the domestic improvements to IPRs outlined above, these international level changes were informed by the technocratic, evidence-based turn in US policymaking honed during the CDP era. They were the culmination of a series of institutional reforms and policies that nourished an expert-led, evidence informed approach built on the centralized coordination of trade policy across several executive branch departments and agencies.

Over most of the 20th century, US trade policy became more greatly concentrated in the executive branch. Constitutionally, Congress held power over tariffs and foreign commerce, but since the Reciprocal Trade Agreements Act (RTAA) of 1934, the legislative branch delegated significant negotiating authority to the President and took on more of an oversight and ratification role.[31]

---

[31] Congress delegated authority to the President to negotiate bilateral tariff-reduction agreements within specified limits. Under the 1934 RTAA, the President could proclaim these rate cuts (up to 50%) without returning to Congress for approval. Decades later, the Trade Act of 1974 introduced the modern "fast track" procedure (now Trade Promotion Authority), which restored a congressional vote but streamlined it: the President gives notice, and Congress must vote up-or-down within strict time limits (e.g., 60 or 90 days) without amendments. This delegation transformed the tariff-setting process by replacing direct legislative tinkering with executive negotiation. Although initially designed as a temporary measure with a three-year sunset provision, the RTAA proved so effective that it was repeatedly renewed in 1937, 1940, and 1945. This framework ultimately became the statutory foundation for US participation in the General Agreement on Tariffs and Trade (GATT) in 1948, establishing a pattern of executive-led trade negotiations that would characterize American trade policy after World War II. This delegation was repeatedly renewed (in 1974, 1994, 2002, and 2015). By the late 1970s and beyond, Congress established broad objectives and consultation requirements, but it was the executive agencies that did the heavy lifting of crafting trade deals and policies in line with those objectives such as market access, job creation, and labor compliance. Once those negotiating objectives were certified, presidents effectively held negotiating authority—subject only to an

A New Approach is Born

Beginning with Carter, the White House orchestrated inter-agency coordination under a data-driven approach that pursued greater liberalization (USTR 2025). Through EO 12188, Carter rationalized the bureaucracy by elevating the Office of the US Trade Representative (USTR) into the nation's primary negotiator, while leaving technical enforcement tasks—such as antidumping and tariff collection—to the Department of Commerce (DoC) and the International Trade Commission (ITC) (see Akhtar 2013). To bind these disparate arms together, the order strengthened the Trade Policy Committee (TPC). Chaired by the USTR, this Cabinet-level forum included not only the Secretaries of State, Treasury, and Defense, but also the government's top economic experts (the CEA Chairman and OMB Director), ensuring that diplomatic strategy was anchored in rigorous economic analysis (see DoS 1980).

This mandate found its operational home in the TPC and its working-level arm, the Trade Policy Staff Committee, which served as the crucible where competing agency views were reconciled. By forcing diplomats, economists, and defense officials to debate their interests before reaching the Oval Office, this structure allowed the executive branch to present a unified position. Importantly, this consensus was built on a foundation of technical analysis: the framework ensured that rigorous evidence from the various agencies flowed upward, filtering out weak proposals before they ever reached the President.

Carter's USTR set a sterling standard that future trade representatives would emulate, as the new office marshaled analytical inputs from Commerce's trade bureaus, the nonpartisan ITC,

---

up-or-down congressional vote—to bind the country to cutting or altering U.S. tariff schedules and other trade rules (see Irwin 2017 on all these points).

and the CEA, and put forth economic analyses that quantified how tariff cuts and similar measures would boost real income and lower consumer prices (CEA 1979; ITC 1978). During the Tokyo Round of the General Agreement on Tariffs and Trade (GATT), negotiated between 1973 and 1979, USTR cited ITC studies and CEA forecasts to press for deeper industrial-tariff reductions and new rules on subsidies and import licensing (USTR 2025; ITC 1978).

In this vein, in a message to Congress explaining the Trade Agreements Act of 1979, Carter explicitly framed these commitments in technocratic terms: as an "anti-inflation" measure that would assure American consumers and businesses access to lower-cost imports and help stabilize prices (Carter 1979).[32]

Next up was President Reagan, who made trade policy a core expression of his free-market agenda. In 1981, the White House issued a "Statement on U.S. Trade Policy" that, for the first time, explicitly endorsed "free trade" as an objective crucial to economic strength and pledged to "strongly resist protectionism," warning that foreign subsidies and market-distorting interventions—not imports themselves—posed the real threat to U.S. competitiveness (see

[32] I should also note that Carter reorganized the DoC in 1979–80 to bolster its trade expertise. He transferred key trade functions (like administering anti-dumping and countervailing duty laws) from Treasury to a new International Trade Administration (ITA) inside Commerce. This placed specialist economists and trade analysts in charge of trade remedy enforcement, aligning decisions with technical findings (e.g. calculating "fair value" prices or foreign subsidy margins) (Casey 2024). When protective actions were taken, they were implemented through rule-bound processes and were anchored by empirical evidence (ibid).

Reagan 1981b; Richman 1988). He bolstered this rhetoric with historical references to the virtues of open global markets (Reagan 1982; Reagan 1986).

This agenda transcended mere ideology: throughout the 1980s, when Reagan pursued free trade, rigorous economic analysis did not trail far behind. Guided by agency projections that lowering barriers would fuel growth, negotiators hit the ground running when the Uruguay Round of multilateral trade liberalization was launched in September 1986. They also aimed to link IP to these discussions; a strategy later validated by analyses projecting significant GDP gains from the final package because of this move (GAO 1994; ITC 1994).

Moreover, Reagan was furnished with the institutional tools to advance a free trade agenda. Congress continued to provide "fast track" authority in 1984, enabling implementing legislation for major trade agreements to receive expedited consideration that would culminate in a simple up-or-down vote (see WaMC 2015). Later, the Omnibus Trade and Competitiveness Act of 1988 institutionalized a fact-based enforcement process. It required the USTR to identify, on an annual basis, countries that denied adequate and effective IP protection or fair and equitable market access and linked "Priority Foreign Country" designations to accelerated Section 301 investigations— a trade remedy provision that empowered the executive branch to investigate and sanction foreign practices that violated trade agreements or burdened US commerce—that could result in retaliation (US Congress 1988; USTR 1989).

While the Reagan administration had already demonstrated its resolve—in the late 1980s, Section 301 was used against Brazil in a pharmaceutical-patent dispute, culminating in tariff retaliation in 1988 (Sell 2003)—the first Special 301 Report (1989) placed Brazil and India on

the Priority Watch List, while Argentina and Egypt were placed on the Watch List.[33] By sowing the fear that escalation to "Priority Foreign Country" status would translate into punitive sanctions, this muscular approach achieved its intended result: targeted nations—including South Korea and Taiwan—strengthened their IP laws to avoid retaliation (Bayard and Elliott 1994).

Clinton Supercharged the Approach

By creating new White House structures that marshaled economic expertise, President Clinton further cemented this evidence-based, executive-led approach to trade and IP policy in the 1990s. In 1993, EO 12835 established the National Economic Council (NEC), chaired by the President, to coordinate domestic and international economic policymaking (Clinton 1993; Clinton White House 1995). Contemporary accounts describe the NEC as a White House–based "honest broker" designed to structure interagency deliberation on economic issues including trade. Staffed with specialists drawn from Treasury, State, Commerce, and the USTR, the NEC ensured that every presidential trade initiative was vetted through rigorous analysis (Wartell 2008).

At the same time, Clinton relied on his CEA workhorse—an independent body of academic economists charged with offering expert, data-driven guidance-to validate those findings (CEA 1996). When the administration pushed the NAFTA through Congress in 1993–94, it leaned on ITC studies forecasting net job gains and consumer savings (ITC 1993). CEA economists supplemented those projections with their own econometric modeling, quantifying expected export growth and wage benefits to counter critics' warnings of massive job losses

[33] Section 301 was also used in the 1980s to address semiconductor dumping by Japan, leading to a landmark U.S.-Japan agreement in 1986 that opened Japanese markets to U.S. firms and set a precedent for future high-tech trade disputes (Irwin 1996; Reagan Library 1986).

(CEA 1996; ITC 1993). Likewise, as the Uruguay Round negotiations wrapped up in 1994 and the WTO emerged, the NEC convened the USTR, Commerce, Treasury, and CEA staff to review a comprehensive ITC impact analysis (ITC 1993).

Impact of an Evidence-Based Approach on Tariffs and IPRs

From the late 1970s onward, US tariff levels declined substantially, alongside successive multilateral negotiations and preferential trade agreements. This trend was anchored by an evidence-driven, executive-led approach that utilized numerous rounds of international negotiations under the GATT framework along with signing onto strategic bilateral deals.

The Tokyo Round GATT agreements, implemented in 1980, represented a major multilateral effort in which representatives of ninety-nine nations worked for five years to remove thousands of tariff and non-tariff barriers (Carter 1979; CBO 1979). Spearheaded by USTR Robert Strauss, this round produced new "codes" —voluntary, plurilateral agreements— that targeted non-tariff barriers by establishing uniform international rules governing subsidies, antidumping, standards, and government procurement (Strauss 1979). The resulting Trade Agreements Act of 1979 (see US Congress 1979b) slashed industrial tariffs by roughly 30 percent—phased in over eight years—and revised domestic procedures to include a "material injury" test for countervailing duty determinations (Carter 1980c).[34]

[34] The "material injury" test was a significant concession in the Tokyo Round. It required U.S. domestic industries to prove that they were being "materially injured" by subsidized imports before countervailing duties could be imposed, bringing U.S. law into closer alignment with international GATT standards.

The subsequent Uruguay Round (1986–1994) expanded this liberalization into agriculture, services, and intellectual property. Take the applied, weighted-mean tariff rate series: the US rate fell from about 3.9% around 1990 to about 3.75% in 1993, continuing downward to approximately 1.55% by 2007 (see Yale Budget Lab 2025).[35] This steep reduction in import duties was informed by sectoral studies and ITC analyses that extolled the mutual benefits of barrier reductions (CBO 1979).

The scope of trade agreements also expanded during this period. Prior to the late 1970s, the US had no free trade agreements (FTAs) outside GATT. The executive branch subsequently negotiated the landmark US–Israel FTA, which entered into force in 1985 (USTR 1985). This was followed by the bilateral agreement with Canada in 1988, which eventually grew into NAFTA when Mexico joined in 1994, creating one of the world's largest trade blocs by GDP (Villareal 2017).

NAFTA exemplified executive-centric, evidence-backed trade policy. Negotiations involved technical working groups and ITC sector studies to ground US positions in projected economic impacts (ITC 1993). Regarding agriculture, Mexico replaced its import licensing requirements with tariff-rate quotas or ordinary tariffs, with 15-year phase-outs established for sensitive commodities like corn and dry beans. This analysis provided an official prospective assessment that shaped the information environment for congressional consideration. Ultimately,

[35] US tariff levels can be measured by several metrics. The "applied, weighted-mean" rate is generally considered the most accurate reflection of the actual costs faced by importers, as it weights the tariff by the value of the goods that exchange hands rather than just averaging every line in the tariff schedule (see WTO 2021).

the Clinton administration used this evidence-based framework to build the bipartisan support necessary for NAFTA's passage, with the House approving the act 234–200 and the Senate by 61–38 (Gerstenzang 1993).

Preeminence of Stronger IPRs

Alongside efforts to increase freer trade, the US executive branch secured substantive international protections for IPRs during this period, primarily through the WTO framework. While earlier agreements like the Paris Convention (1883) established the principle of national treatment—that a country should treat foreign creators the same as its own (see PCPIP 1883)—they lacked effective enforcement mechanisms such as the retaliatory trade leverage provided by the WTO's Dispute Settlement Understanding (Long 1997).

Conversely, the Agreement on TRIPS, adopted in 1994 and entered into force in 1995, fundamentally rewrote the international IP landscape. It established substantive minimum standards across all IP domains, requiring member states to grant patents for inventions in all fields of technology for a minimum term of 20 years, while extending protection to computer programs and integrated circuit layouts (WTO n.d.a). By linking minimum IP standards to the WTO's binding dispute-settlement system, TRIPS made IP violations subject to the same cross-sectoral trade sanctions used for disputes over physical goods. It mandated that members make "effective" enforcement procedures available, including the authority for injunctions, the seizure of infringing goods, and border measures allowing customs authorities to suspend the release of infringing imports (see WTO n.d.b).

NAFTA achieved similar ends with similar means. US-led technical teams drafted provisions to align Mexican patent laws with US standards, including protections for pharmaceuticals and biotechnology (see Villareal 2017). These measures responded to industry

concerns that Mexico's weak IP enforcement cost them billions of dollars in lost revenue (USTR 1993).

Federal Encouragement of Standard Essential Patents

We've now reached the point where I will move to the last entry in Table 3.2, the final area where the US government bolstered stronger IPRs during the CDP era. By encouraging federal agencies to participate in private standards bodies, the Executive Branch empowered SSOs, endowing them with tools that helped industry leaders make digital devices interoperable.

A Quick Primer on Standard Setting Organizations

SSOs coordinate technical standards and typically require participants to disclose potentially essential intellectual property and to make FRAND (fair, reasonable, and non-discriminatory) licensing undertakings.[36] Because licensing terms are generally negotiated bilaterally between firms or via patent pools outside the standards body, and because the government does not dictate how this happens, the system relies on "private ordering" to ensure all players can access technology (see Barnett 2020; Haber 2016). Disputes over IP licensing are typically resolved through courts or private arbitration, rather than within the SSO itself (Contreras 2015; Farrell and Simcoe 2012).

The standards under discussion provide architectural certainty that allow platform providers to manage end-user interactions and sustain ecosystems where multiple firms can build interoperable products. By codifying technical specifications—from the physical pin

[36] While the SSOs themselves do not typically "designate" SEPs or negotiate commercial terms, their disclosure policies ensure that potentially standard-essential patents are identified early (see Lemley and Melamed 2013; ETSI 2024).

configurations of a USB-C port (maintained by the USB-IF) to the handshake protocols of Wi-Fi (based on the IEEE 802.11 family)—SSOs ensure that a manufacturer's hardware is compatible with a device's global ecosystem (see Farrell and Simcoe 2012; Barnett 2020). In particular, when it comes to smartphones, 3GPP enables fabless chip firms, handset makers, and infrastructure vendors to collaborate on air-interface standards, which ensures that every component—from a baseband processor to the modem firmware that operates it—conforms to the same technical protocols (Barnett 2020; 3GPP 2024).[37] The end result is a modular smartphone ecosystem.[38]

The Government's Role Facilitating Standard Setting

While SSO participation is purely voluntary and largely self-enforcing, these organizations' effectiveness greatly depends on complementary government institutions and policies. Left to the

[37] The baseband processor is a specialized integrated circuit designed to manage all radio communication functions—such as signal modulation and encoding—required for a device to interface with cellular networks. This hardware is controlled by modem firmware, the low-level software embedded in the chip that executes the complex protocols mandated by SSOs like 3GPP to ensure the device remains synchronized with the network's technical specifications (see Holma and Toskala 2012).

[38] As I will elucidate in Chapter 6, LTE, a prominent standard from 3GPP, significantly boosts mobile data capacity and speed by utilizing an advanced radio interface and enhanced core network infrastructure. Other digital device technologies governed by SSOs include Bluetooth (Bluetooth Special Interest Group), Wi-Fi (IEEE 802.11 family and Wi-Fi Alliance), GPS (GNSS), and fingerprint scanners (FIDO Alliance).

market alone, the collaborative standard-setting process would be paralyzed by the threat of antitrust litigation among horizontal competitors, high transaction costs in identifying essential intellectual property, and patent disputes that potentially derail industry-wide adoption. During the CDP era the federal government actively lowered these barriers, providing critical legal safe harbors and administrative directives. Through those efforts, executive branch agencies like the National Institute of Standards and Technology (NIST), the DoJ, and the FTC played a pivotal role in spurring, legitimizing, and protecting these voluntary collaborative efforts.

First and foremost, SSOs call on industry rivals to literally sit at the same table when creating and negotiating technical standards, a taboo forbidden by antitrust law. Therefore, it took concerted efforts by government agencies to clarify that collaborative activities around standard setting were generally pro-competitive and would not inherently violate antitrust laws.[39]

Consider some examples. The National Cooperative Research Act (NCRA), passed in 1984, was designed to encourage American companies to collaborate on R&D and production joint ventures by, among other things, limiting their exposure to antitrust litigation (US Congress 1984b). Later, the Standards Development Organization Advancement Act of 2004 extended the NCRA protections to SSOs, ensuring that standard-setting is evaluated under the antitrust Rule of Reason rather than be considered illegal per se (US Congress 2004; DoJ 2004). The National Technology Transfer and Advancement Act directs federal agencies to use and participate in voluntary consensus standards wherever possible (US Congress 1996). OMB Circular A-119

[39] As Chapter 7 makes clear, these types of collaborative agreements between competitors usually arouse the suspicion of antitrust authorities that illegal collusion is occurring under Section 1 of the Sherman Act.

codified a federal strategy for standards development that urged agencies to actively participate in SSO efforts and adopt private-sector standards in regulations (see OMB 1998).

The most important vehicle for advancing these objectives is NIST, however. In the case of 3GPP and other SSOs involved in promoting cellular standards, this storied federal agency played a multifaceted role in supporting the wireless and microelectronics standards ecosystem.[40] NIST labs provided the metrological foundations and critical expertise for setting standards related to chip performance and interoperability; NIST itself defined and verified performance standards for wireless signals and component characteristics; and its research groups contributed to technical specifications within SSOs like 3GPP, covering areas like spectrum sharing and 4G technologies.

Beyond technical contributions, NIST served as a neutral platform to convene diverse stakeholders to facilitate consensus building. Behaving as a neutral arbiter helped mitigate risk, since fear of picking the "wrong" standard often leads to market paralysis (see Mazzucato 2013). By hosting widely attended workshops and establishing "testbeds," NIST allowed competing firms to validate their technologies against emerging benchmarks in a pre-competitive environment, effectively reducing the odds that any single firm's proprietary path would lead to a technical dead-end.

[40] This extended passage on NIST draws on scores of its SP-800 Series documents (see NIST 2024) and particularly Golmie et al. (2023) and Bohn et al. (2023).

Federal agencies also actively encouraged and reinforced FRAND licensing during the CDP era.[41] The FTC and DoJ regularly submitted amicus briefs to reassure industry that collaborative standards development would not be viewed as problematic. They also pursued enforcement actions against companies that attempted to manipulate FRAND arrangements through non-disclosure (see Farrell et al. 2007).

The FTC's lawsuit against Dell Computers marked a landmark intervention against "patent ambush," which occurs when an SSO participant intentionally conceals ownership of a relevant patent during standard-setting discussions (FTC 1996). While Dell had actively participated in deliberations held by the Video Electronics Standards Association (VESA) in 1992 on a Vesa Local Bus (VL-bus) standard for PCs, and had certified it held no relevant patents, things changed once the technology went live.[42] After the industry committed to the new standard, Dell attempted to enforce its undisclosed patent; the FTC's resulting consent order compelled Dell to cease and desist from enforcing its patent rights against any computer manufacturer using the VL-bus standard.

---

[41] See Contreras (2022) on the points contained in this paragraph. I should also note that during the CDP era, the federal government also supported the formation and operation of patent pools—such as those for MPEG and DVD technologies—explicitly conditioned on FRAND or similar licensing terms; efforts that, although also helpful to promoting innovation in general and the digital and AI revolutions in particular, are beyond the scope of this chapter (see DoJ 2020).

[42] The VL-bus was a standard for a high-speed expansion slot on a computer's motherboard that primarily connected graphics cards directly to a PC's main processor (see FTC 1996).

The FTC's ruling set an important precedent: it established that a patent holder's failure to disclose intellectual property during a standard-setting process could constitute an "unfair method of competition," effectively transforming private SSO disclosure rules into enforceable public obligations (see Contreras 2015; Barnett 2020).[43]

While these state-enabled coordination mechanisms helped bring about device interoperability in ways that created new markets, they also stoked economic warfare. While every player should prefer a shared standard instead of fragmentation, they remain locked in a distributional struggle over which specific standard—and whose IP—will prevail. By mandating that essential technologies be shared under FRAND terms, Washington, D.C. did not eliminate conflict over how the "producer surplus" is divided among the various layers of global supply chains for digital devices—it brought that conflict into sharper relief (see Galetovic et al. 2024).

Distributional Issues in the Shadow of Innovation

During the CDP era, distributional concerns often shaped policy as much as considerations about dynamic efficiency. Consider the political economy of the vertically disintegrated supply chains associated with digital devices, a topic I explore throughout this book. By definition, a voluntary exchange between those who generate the ideas that conceptualize or improve these products and those who rent or use them to bring these goods to life is a pie expanding interaction.

However, this does not mean distributional conflicts over the producer surplus magically disappear once these gizmos reach their consumers (Haber and Lamoreaux 2021; Galetovic et al.

[43] However, while the FTC brought a similar suit against Rambus (see FTC 2002), the company ultimately prevailed on appeal when the D.C. Circuit found insufficient evidence of antitrust violations (see Rambus 2008).

2024; Menaldo and Wittstock 2021). Firms at the pinnacle of these supply chains, designers, have strong incentives to fight for stronger intellectual property rights to enhance their negotiating leverage and appropriate a larger share of the surplus. Firms further down the chain, implementers, have equally strong incentives to fight to weaken upstream designers' IPRs, as their share of the profits increases when they negotiate the royalties they pay for these ideas down.

The upshot is that fights over the distribution of the surplus are articulated by antagonists in the language of efficiency: the size of the pie. For example, implementers often accuse designers of raising prices and stifling innovation, arguing that, if left to their own devices, designers will exploit the "market power" conferred by their SSOs and engage in "patent holdup" and "royalty stacking." The idea is that once a SSO incorporates patented technology into a standard and implementers make irreversible investments, patentholders will seek to extract higher royalties; moreover, when a standard relies on multiple patents held by different owners, implementers will be burdened with "excessive" cumulative royalties (Lemley and Shapiro 2007; Shapiro 2010a). In response, designers often downplay the value implementers add to the final product, disparaging their contributions (see, e.g., Sidak 2013).

By extension, both types of players try to lobby policymakers to use the law and regulation to weaken each other's property rights and thus reduce their respective share of the surplus (see Haber and Lamoreaux 2021). For example, in 2021, a coalition funded by Apple and other tech firms urged the Biden administration to reverse Trump-era policies favoring SEP holders, arguing that they harmed small businesses and stifled competition (ACT 2021). Meanwhile, the Innovation Alliance, representing SEP-intensive firms like Qualcomm, defended

patentholders' IPRs as critical to recouping R&D investments, asserting that weakened patent rights disincentivize contributions to standards (see Innovation Alliance 2022).

Implementers Win the Day

During the early-2010s, implementers seized upon these ideas to reach the higher ground. US agencies increasingly accepted the patent skeptic view that, when a SEP is encumbered by a FRAND commitment, patentholders should be restricted from seeking injunctive or exclusionary relief against a willing licensee.[44] In 2012, as a condition of clearing Bosch's acquisition of SPX Service Solutions, the FTC required Bosch to license its SEPs on FRAND terms and to refrain from seeking injunctions against licensees who were willing to pay those terms (FTC 2012). In 2013, it decreed that Google/Motorola had to offer binding arbitration or accept a court's FRAND determination before pursuing exclusionary relief, and it could proceed only if an implementer was unwilling to take a neutral FRAND license (FTC 2013a). That same year, the DOJ and USPTO jointly warned that when a patentholder has pledged to honor FRAND licensing terms, an injunction or exclusionary order to enforce an SEP may be inconsistent with the public interest "in some circumstances" (DOJ and USPTO 2013: 5).

[44] While both remedies serve to stop patent infringement, they involve different "levers" of state power. An injunction is a judicial order issued by a federal court that commands a specific company to stop its infringing activity; failure to comply is treated as a direct challenge to the court's authority. By contrast, an exclusionary order is an administrative tool used by the International Trade Commission (ITC) to protect domestic markets. Rather than targeting a company's overall behavior, it targets the physical products themselves by authorizing U.S. Customs to seize or block them at the border.

Importantly, in 2015, the DoJ issued a favorable business review of the IEEE's updated patent policy—a move widely viewed by scholars and industry participants as a federal endorsement of a "patent-skeptic" agenda (see Barnett 2020; Contreras 2017). This policy included provisions restricting the ability of SEP holders to seek injunctive relief against willing licensees and clarified that reasonable royalties should be based on the smallest salable patent-practicing unit. By sanctioning these changes, the DoJ effectively signaled that royalties should be calculated based on the price of the specific component that implements the technology, such as a baseband chip, rather than the price of the entire end-user device like an iPhone (DoJ 2015).

Designers Strike Back

However, buoyed by clearer thinking about these ideas and new evidence, designers began to regain lost ground during the late 2010s. Drawing on Williamson's (1985) view of opportunistic holdup, Galetovic and Haber (2017) argue that different players in a supply chain rightly anticipate the possibility that once they make fixed, irreversible, and asset-specific investments, they may threaten each other's quasi-rents—the difference between their short-run and long-run marginal costs. Therefore, rather than relying on regulation and litigation ex post, they tend to adjust their relationships and behavior to avoid holdup ex ante.[45] Moreover, empirical evidence

[45] The term "quasi-rent" carries two related meanings in the transaction-cost literature. In the narrower formulation associated with Klein et al. (1978), quasi-rent refers to the difference between an asset's return in its current specialized use and its return in the next-best alternative deployment—the gap that becomes appropriable precisely because relationship-specific investments cannot easily be redeployed. In the broader formulation employed by Galetovic and Haber (2017), quasi-rent denotes the difference between a firm's total revenues and its short-run

does not support the view that SEP owners routinely exploit injunctions to extract excessive royalties or that royalty stacking is a significant problem in practice (see Galetovic et al. 2018).[46]

Informed by this and other academic work, in 2019 the DoJ, USPTO, and NIST espoused the view that SEPs should not be treated differently from other patents and that all remedies (including injunctions) should be available when appropriate (DoJ/USPTO/NIST 2019).[47] They also cautioned that depriving patent owners of injunctive relief could harm innovation by encouraging "hold-out"—if implementers know they can only be sued for a "reasonable royalty," they might delay signing a license. Finally, these agencies also pointedly noted that implementers

---

variable costs—the maximum amount extractable before the firm hits its shutdown condition. These definitions converge in practice: asset specificity implies low salvage value, which in turn means the firm's outside option yields little revenue, widening the gap between current earnings and the shutdown point. Both formulations identify the same underlying vulnerability: once investments are sunk, the investing party can be squeezed down to the point where continued operation barely covers variable costs, with no recovery of fixed capital outlays.

[46] In Chapter 6, I argue that SEPs are not foolproof instruments for upstream firms to recover their long run costs so that they can perpetually innovate—IPRs are complemented by alternative strategies such as diversifying the base of implementers (e.g., digital device manufacturers).

[47] However, they reaffirmed that any specific FRAND commitments made to an SSO must be considered in court decisions. This signaled a return to a more neutral, case-by-case approach grounded in general patent and contract law using the default eBay standard for injunctions that recognized FRAND as a contractual limitation (see Kessler 2022).

could potentially exercise monopsony power by colluding to suppress royalty prices—signaling that anticompetitive behavior by licensees is also a concern.

Alas, the melodrama did not end there. The FTC did not join the DoJ and NIST joint statement because, at the time, it was pursuing an SEP-related lawsuit against Qualcomm for charging "unreasonable royalties"—a case it eventually lost when the Ninth Circuit Court found no evidence of anticompetitive holdup (see Qualcomm 2020). Then, in 2022 the Biden administration withdrew the 2019 pro-SEP policy (see DoJ/USPTO/NIST 2022)—only to have the second Trump administration reverse course yet again.

Rather than issue a public decree, the White House pursued surgical interventions this time around. The USPTO spearheaded efforts to make it easier for inventors to block infringers by filing "statements of interest" in court. In one of those cases, it teamed up with the DoJ to pointedly argue that a patent is a unique asset, and if a court only requires an infringer to pay a small fee rather than stopping the use of the invention, it effectively strips the inventor of their right to control their own property (DoJ and USPTO 2025). Furthermore, the DOJ and USPTO urged the ITC to treat blocking orders as the "default" response to infringement, arguing that the government should stop treating the public's desire for "cheap products" as a valid excuse to let companies get away with using stolen technology (DoJ/USPTO 2025).

No Matter, the Result Was Innovation

Notwithstanding these seemingly drastic swings in perspective from one administration to another, we should not miss the forest for the trees: the presence of clear FRAND commitments, backed by predictable SSO policies and the government's general support, was critical to enabling collaborative innovation in the run up to the Fourth Industrial Revolution. A company developing a new smartphone or Wi-Fi device during the CDP era that relied on dozens of

interoperable standards could easily license the needed SEPs on reasonable terms. Likewise, inventors who contributed technology to a standard knew they could obtain a fair return on their investment. On balance, the overall winners were investors, consumers, and the public at large, even if policy seesawed back and forth and sometimes modified the split of the surplus enjoyed by these parties.

OTHER CDP REGULATIONS

Creative Destruction Paradigm era policymakers didn't just help lower production costs, they also indirectly reduced coordination and contracting costs, making entirely new categories of transactions possible. Specifically, new regulations around electronic contracting, authentication, and platform-mediated exchange routine enhanced legal certainty and standardized procedures, facilitating more efficient and secure online transactions.[48]

The Uniform Electronic Transactions Act (UETA), adopted in 1999 (see ULC 1999) and the Electronic Signatures in Global and National Commerce Act (E-Sign), adopted in 2000 (see US Congress 2000), revolutionized digital commerce by establishing the legal validity of electronic signatures and records equivalent to paper documents. Key court decisions reinforced this legal foundation: *Specht et al.* (2002) established the requirement for conspicuous terms and

[48] While in what follows I will discuss legal reforms that have reduced certain transaction costs by standardizing and validating electronic documents and transactions, I should note that persistent challenges betray the promise of a seamless, integrated global digital network; these include the inconsistent adoption of a digital approach to contracting and commerce, disparate regulations at the subnational and international level, the growing costs of digital fraud, and ongoing cybersecurity concerns persist (see Gibson 2015; Nagle et al. Seamans 2020).

clear user assent, while *Barnes & Noble* (2014) confirmed that users must have actual or constructive notice for terms to be binding (see Mozingo 2020).

On the one hand, they standardized electronic transaction laws across the US, significantly reducing barriers to digital business, and providing a more predictable legal environment (Stern 2001; Gibson 2015). On the other hand, they bolstered consumer confidence in e-commerce (see Geva 2020). In turn, these laws supercharged digital interstate commerce (see Gibson 2015).

Similarly, privacy and data protection regulations, the legitimization of electronic records, and the advent of online dispute resolution mechanisms all contributed to increased trust in digital transactions, enabling businesses to operate more efficiently and expand their footprint (Schwartz and Janger 2002). Privacy and data protection regulations, such as the Gramm-Leach-Bliley Act and FTC guidelines (US Congress 1999; FTC 2000), increased user and business confidence in online transactions and bolstered e-commerce (Schwarts and Janger 2002). Electronic record-keeping laws, including the E-SIGN Act and UETA, recognized electronic records as legal documents in business and government transactions (see US Congress 2001), reducing the need for paper records, lowering storage and retrieval costs, and making it easier for private and public actors to manage and exchange information digitally (Richards 2025; NCCUSL 1999). Support for online alternative dispute resolution mechanisms provided cost-effective means for buyers and sellers to resolve issues without resorting to costly litigation.[49]

[49] While international agreements and initiatives made progress toward harmonizing e-commerce and digital trade laws, significant differences and regulatory fragmentation remained—especially regarding data privacy, taxation, and consumer protection, a topic I will return to in Chapter 12.

CONCLUSION

This chapter has traced how, over the final decades of the twentieth century, the United States shed its earlier populist-statist instincts and coalesced—seldom along strict ideological lines—around a Creative-Destruction Paradigm that married rigorous, evidence-based analysis with a muscular, yet targeted, role for government in shaping markets and fostering innovation. We saw how, under successive administrations, formal institutions and legal reforms—from the APA through EO 12866 and Circular A-4—embedded cost–benefit reasoning into rulemaking; how the Court of Appeals for the Federal Circuit, the Bayh–Dole Act, software-patent jurisprudence, and international agreements such as TRIPS realigned property-rights incentives; and how voluntary SSOs and interagency trade-policy coordination solved classic coordination, distributional, and commitment problems to unlock dynamic efficiency.

Equally important, this chapter has shown that these reforms were not simply "neoliberal" retreats from state action but rather a bipartisan, technocratic project to harness government's capacity to create focal pivots, fair-split rules, and credible commitments. By assigning and enforcing a new set of property rights, reducing transaction costs associated with standardization, coordination, negotiation, and contract enforcement, and solving an array of market failures, the CDP profoundly transformed America's political economy.

As the next chapter will show, the Creative Destruction Paradigm gave birth to new networks of firms contracting with each other and drastically changed supply chains and business strategies. As a result, new markets, firms, patterns of industrial organization, business models, products, and digital services proliferated. In a virtuous circle, these innovations further reduced transaction costs.

CHAPTER 4. SUPERSTAR DYNAMICS

Using a range of evidence from trade statistics to the changing composition of corporate assets to patenting and patent citation data, this chapter will detail an epic economic transformation. It will tell the story of how the U.S. economy evolved from a traditional manufacturing model centered on heavy industry to one dominated by intangible assets like IP, software, and brand equity. This dynamic was driven by the vertical disintegration of global supply chains, a process where American firms specialized in high-value design and R&D while outsourcing physical production and exporting high-value services like finance, software, and consulting.

Furthermore, this chapter will review how this transformation marked some US companies as "superstars". These leaders were able to separate from the pack by leveraging unique intangible assets, unprecedented scale, and mastery over supply chains. Greater productivity meant higher profits.[50] This was particularly the case for high-tech firms focused on industries such as semiconductors, computer manufacturing, software publishing, and telecommunications, and especially for those operating within the “digital stack.” Together, they helped create the infrastructure upon which big datasets, machine learning, artificial intelligence, and the internet of things would flourish.

The next two chapters will expand upon this story, revealing how elite, pacesetting technology leaders considerably outraced their rivals in inventive capacity and influence over technical standards and best practices. But before we detail how superstars at the “head of the pack” dictated the pace of technological change, I focus our attention on the bigger picture.

---

[50] A large literature corroborates these insights and includes Autor et al. 2020; Bessen 2020; Galetovic 2021; Gochberg and Menaldo 2022; and Crouzet and Eberly 2019.

GLOBALIZATION's MARCH

The immediate aftermath of World War II launched an era of renewed globalization.[51] This age witnessed unprecedented economic integration, marked by an extraordinary increase in international trade, capital flows in the form of cross-border lending and portfolio flows, and technological exchanges (see Menaldo and Wittstock 2021). It also saw the expansion of global supply chains that let the United States specialize in IP- and services-intensive activities. Global integration became so extensive that disruptions in one part of the world rapidly reverberated across international economic networks, as exemplified by the supply side shocks induced by the COVID-19 pandemic in 2020 (Lipscy 2020).

During the first phase, immediately after the end of the war, Bretton Woods Institutions like the International Monetary Fund (IMF) and World Bank helped stabilize trade and financial

[51] An earlier "first wave" of globalization that spanned roughly between 1870 and 1914 was underpinned by the classical gold standard, falling transport and communication costs (associated with the diffusion of steamships and the telegraph, and London-centered finance. That era saw substantial integration in goods, capital, and labor markets—evident in price convergence, large cross-border portfolio flows, and mass migration (O'Rourke and Williamson 1999; Findlay and O'Rourke 2007). It unraveled with World War I and was not reconstituted during the interwar years: efforts to restore gold at misaligned parities transmitted deflation; the Great Depression triggered tariffs and quotas (e.g., Smoot-Hawley) alongside exchange controls, clearing arrangements, and imperial preferences; international lending dried up as multilateralism fractured into blocs (Eichengreen 1992).

flows and rebuild war-torn economies.[52] The first, a global lender of last resort, lends money to countries struggling with fiscal crises that are unable to finance their budget deficits. The second, a development bank, offers subsidized loans to developing countries to finance infrastructure and similar investments.[53]

The US and its closest allies cultivated trading networks under the aegis of this "Bretton Woods" system.[54] During the 1960s, Western nations accounted for a significant share of global

---

[52] The distinctiveness of the post-1945 era lies less in the mere existence of cross-border exchange than in its breadth, depth, and institutional architecture (Baldwin 2016).

[53] An important backbone of this economic liberal order was hard money in the way of the US monetary regime and financial system: it reduced currency risk, standardized payments and settlements, and lowered funding costs. Right out of the gate, a gold–dollar system hybrid stabilized expectations: other currencies were pegged to the dollar, and the dollar promised convertibility at $35 per ounce. Later, when the U.S. closed the gold window in 1971, globalization did not reverse; rather, it blossomed under the now free floating US dollar, which offered institutional credibility and scale: the unrivaled depth and continuous liquidity of the U.S. Treasury market, open capital accounts, vibrant financial markets and intermediaries, and a predictable rule-of-law framework buttressed by an independent central bank (Rogoff 2025). Network effects locked in the dollar for trade invoicing, commodity pricing, and international intermediation, while Treasuries furnished ultra safe assets.

[54] The system is named after the location of the "United Nations Monetary and Financial Conference," held in July 1944 at the Mount Washington Hotel in Bretton Woods, New

exports (Cohen and Sisle 1970). By 1972, countries in Northwestern Europe and the US still controlled over half of world exports, reflecting their industrial strength and economic integration during this period (Federico and Tena-Junguito 2018). By 1979, one in every seven US manufacturing jobs was export-related (US Census Bureau n.d.). In the US and other developed countries, "embedded liberalism," in the form of welfare states intended to stabilize economies and reduce volatility, moderated increasing global integration and helped legitimize it (Ruggie 1982).

In terms of the rest of the world, there were a lot of heterogeneous experiences with the global economy during the Cold War.[55] The Soviet Union, Eastern bloc countries, and other communist countries like North Korea and Cuba traded amongst each other outside of the Bretton Woods system. Many developing countries in regions that spanned from Latin America to Sub Saharan Africa to the Middle East indulged in protectionism and import substitution industrialization to try to switch from commodity exports to indigenous manufacturing. This group even included Taiwan and South Korea, although they differed from their peers in their efforts to export finished goods such as cars to the industrialized world even if tariff barriers protected key domestic industries from foreign competition.

Post-Cold War Globalization Boom

Hampshire. As World War II ended, delegates from 44 Allied nations met to design a new international economic framework that would promote stability and prevent a repeat of the economic chaos of the 1930s. The agreements reached there led directly to the creation of the IMF and the World Bank.

[55] See Frieden (2006) for the points across both ensuing paragraphs.

As the new millennium approached, and especially in the wake of the Soviet Union's downfall, major economies went about dramatically lowering their import tariffs.[56] Consistent with the story I told in the previous chapter, between 1988 and 2017, the average applied most-favored nation tariff globally was nearly halved, from about 16% to 8.5%—with the steepest declines occurring between 1994 and 2005.[57] By 2016, the United States' trade-weighted average tariff had reached 1.67%, a historic low. Many developing economies followed suit: China's tariffs fell by more than 30 percentage points, India's by over 50, and Brazil's by about 10 (see Bandyopadhyay et al. 2022).[58]

What's more, the infrastructure for containerized seaborne cargo underwent major upgrades during the 1980s and 1990s. These included the emergence of mega-vessels capable of transporting vast quantities of goods; the automation of port operations to speed up handling and reduce costs; and the integration of inland freight networks to ensure seamless movement of raw materials, intermediate inputs, and finished products across borders.

In parallel to all these developments, global trade volumes mushroomed. Measured by volume, world merchandise trade consistently outpaced global output in the late-20th century and

---

[56] These figures are available in Teti (2020).

[57] The most-favored-nation principle, a cornerstone of the WTO, which was birthed by the Uruguay Round itself, requires a country to apply the same tariff rates to all its fellow WTO members.

[58] By the late 1980s and early 1990s, many developing countries began to liberalize their capital accounts as well (see Albertus and Menaldo 2018)—albeit, with delayed or more gradual financial liberalization in countries like Argentina, Brazil, and India.

early 2000s—averaging about 6% annual growth in 1997–2007 versus roughly 3% for GDP. All told, trade volumes grew about 45-fold between 1950 and 2008 (see WTO 2008).[59]

Finally, trade liberalization and capital mobility not only lowered input prices, but also accelerated technology transfer, allowing even less-developed economies to import advanced machinery and know-how via FDI and licensing agreements on the back of stronger IPRs (Menaldo and Wittstock 2021; Romer 1993).

The Increasing Role of Asian Countries

Under Deng Xiaoping's leadership, China adopted market-oriented reforms, including joint ventures and special economic zones, to attract foreign investment and accelerate modernization (Romer 1993). China's liberalization was a gamechanger, allowing it to join the ranks of already established manufacturing powerhouses such as Japan, Taiwan, and South Korea. Consequently, Asia's share of global manufacturing output growing from 26.5% in 1990 to 46.5% in 2013.[60] In 1990, China produced less than 3% of global manufacturing output; by 2015 it was nearly a quarter. That year, China produced 80% of the world's air-conditioners, 70% of its mobile phones, and 60% of its shoes.

---

[59] The post-World War II era of globalization started to peter out after the Global Financial Crisis in 2008, however—with its dénouement accelerating during the first Trump administration, beginning in 2017. The world trade-to-GDP ratio peaked at 61% in 2008. The US trade-to-GDP ratio peaked at 31% in 2011 and US-China trade grew significantly after the 1990s, with imports from China reaching a high of 22% of total U.S. imports in 2017.

[60] For the facts and figures in this passage see The Economist (2015a) and The Economist (2015b).

Despite challenges such as Chinese manufacturing wages rising by more than 10% annually between 2001 and 2015, and the yuan reaching an all-time high in 2015, China boasted dense clusters of suppliers and customers and improving infrastructure that lowered production costs. Chinese labor remained relatively cheaper and more productive compared to rich-world rivals, while investments in automation increased and buoyed productivity gains.

And while companies like Foxconn adapted to higher Chinese wages by moving operations to lower-cost interior regions such as Henan and Sichuan chasing tax breaks and cheaper land, South-East Asia emerged as an important manufacturing hub. Lower wages and younger populations endowed countries in the region with a competitive edge, while proximity to China and rising Chinese consumption benefited regional manufacturing. Garments and low-skill manufacturing moved to countries like Myanmar, while more advanced manufacturing such as vehicle production shifted to Vietnam, Thailand, and Indonesia.

Increased Globalization was Largely a Choice

As I outlined in the previous chapter, the upsurge in globalization that emerged after the post-Cold War came about largely because the U.S. deliberately consolidated an international economic system designed to promote cross-country trade, capital mobility, and technology transfers (see also Rodrik 2011: Ch. 1–2). Indeed, despite the collapse of the WTO's Doha Round of trade liberalization efforts in the mid-2000s—largely over disagreements about reducing agricultural import barriers—a robust bipartisan consensus persisted in the U.S. in favor of deepening economic ties with the rest of the world and especially China, driven in part by the

desire to help American firms globalize their supply chains to reduce costs and to gain access to China's rapidly expanding consumption market (see Pregelj 2001).[61]

Other countries followed in kind for similar reasons. Many western governments embraced a more liberal approach to trade and investment in the belief that by exploiting their comparative advantages in high-value added manufacturing and services, global supply chains and minimal barriers would yield shared prosperity (Frieden 2006: 398–400; Rodrik 2011: Ch. 6). Similarly, developing countries jumped on the same bandwagon to try to gain a slice of a growing global pie—albeit sometimes serious economic crises made it so that they had little choice but to liberalize their economies (Albertus and Menaldo 2018; Haggard and Kaufman 1992).

Overall, the US Benefited Greatly from Globalization

This international economic system really paid off for key American industries that were globally interdependent (see Menaldo and Wittstock 2021). Between 1992 and 2020, high-tech sectors frequently outpaced real GDP growth, doubling or even trebling it: While U.S. real GDP grew at an average annual rate of about 2.5% during this period (BEA 2021), sectors like information technology (IT) averaged growth rates between 5% and 7% annually (National

[61] Scholars have put forth other reasons for why the US pursued this objective: to bolster its political and military hegemony and promote democratization and peace—both directly, through greater integration with the global economy, and indirectly, via the prosperity that would ensue (e.g., Nye 2004: Chapter 3).

Science Board 2020).[62] The global semiconductor industry saw sales grow from approximately $60 billion in 1992 to over $440 billion in 2020, reflecting an annual compound growth rate of 8% (SIA 2021). The software market similarly expanded during this period, rising from around $60 billion to over $530 billion (Mehta et al. 2021).

These "winning" sectors obtained an impressive global reach in the process. U.S.-headquartered semiconductor companies have long maintained a strong global presence, with a significant majority of their sales—often exceeding 80%—going to foreign markets. As we shall see in the next chapter, this high export intensity reflects the industry's deep integration into global supply chains (SIA 2020). By 2023 the US exported over $300 billion in information and communications technology and business services, resulting in a net surplus of $120 billion (BEA 2024). Additionally, U.S. royalty and license fees, primarily from technology, achieved a net surplus of $90 billion in 2022 (World Bank 2023). Financial services also contributed significantly, generating a surplus of $63 billion in 2023 (CEA 2024).

HOW THE US BECAME AN INTANGIBLE CAPITAL ECONOMY

As outlined above, the dollar–treasury and international trade architecture imposed and sustained by the US contributed to cross-border trade and investment flows. It did so by significantly lowering transaction costs in terms of strengthening property rights and international contracting and reducing currency risk and standardizing settlements.

[62] It is important to note, however, that R&D-intensive, intangible-capital-rich industries grew at a rate that was twice that of real GDP between 1972 and 1992 and their export shares were also significantly higher than the national average during this earlier period (see FTC 1996).

I now show how that scaffolding reshaped the *structure* of U.S. trade. While at first this was achieved by deepening vertical disintegration in manufacturing, it later centered much more on shifting export earnings toward services and intangibles.

Vertical Disintegration

Until the late 20th century, large firms—from Rockefeller's Standard Oil to IBM, AT&T, and Xerox—tended to own every step of their value chains: they controlled raw-materials extraction, component fabrication, final assembly, distribution, and after-sales service, almost always within their home countries (see Menaldo and Wittstock 2025). But by the 1990s, advances in ITC technologies, the liberalization of trade and capital flows, and stronger international IP regimes (see Menaldo and Wittstock 2021) combined to make outsourcing not only feasible but irresistibly efficient.

The first wave of horizontal specialization took hold well before 2000. Japanese and Korean electronics makers assembled PCs around American-designed microprocessors and operating systems (O'Mara 2019; Miller 2022), while US automakers such as Ford and General Motors opened plants in Mexico and Canada—anticipating NAFTA—to tap lower-cost labor and sidestep trade barriers. Around the same time, Eastman Kodak handed its data-center operations to IBM, General Electric sent support jobs to India, and Nike relocated nearly all shoe production to factories in Asia, retaining only design, marketing, and global sales at its US headquarters (see Baldwin 2016).

By the early 2000s, this playbook had become universal. For example, Xerox outsourced its entire IT infrastructure to Electronic Data Systems (EDS) in a landmark 10-year, $3.2 billion deal in 1994 (Applegate and Davis 1995), Procter & Gamble farmed out datacenter and network management to HP in a $3 billion, 10-year agreement in 2003 (Junarkar 2003), and Ford spun off

its parts division as Visteon in 2000 to focus on core vehicle design and assembly. In aerospace, Boeing outsourced over 70 percent of the 787 Dreamliner's value chain—everything from engine casings to wing sections—to a web of global suppliers, consolidating its own role as integrator rather than fabricator. (see Peterson 2011).

These corporate decisions, repeated across industries, created expansive, vertically disintegrated ecosystems in which specialized firms could invest intensely in their niches and achieve global scale (Baldwin 2016). The result of all this outsourcing and globalization was unprecedented efficiency: each segment of different supply chains concentrated on its core competencies, also benefitting from global labor markets (for instance, India's software engineers) and specialized spatial clusters (see Menaldo and Wittstock 2025). And because a robust international IP regime reduced the risk of expropriation, firms could safely license or outsource critical technologies across borders (Menaldo and Wittstock 2021). In this environment, downstream assemblers competed on razor-thin margins, while upstream innovators captured outsized returns on intangible capital—patents, brand equity, data, and proprietary platforms—laying the groundwork for the "superstar firm" phenomenon I explore further below.

As a way of visualizing how this process unfolded over time, consider Figure 4.1, which depicts the foreign-value-added (FVA) share of US manufacturing exports from 1995 through 2020. For the initial, 1995 to 2010 period, the figure represents an intuitive proxy for vertical disintegration at the global level for a relatively tangible economy with a still visible physical footprint: the rising share of foreign-value-added share in US manufacturing exports reflects the depth of cross-border fragmentation in traditional manufacturing.

As the early fruits of the CDP changes to previously populist-statist legislation and regulation fostered global trade liberalization and the nascent strengthening of IPR regimes (see Chapter 3), US manufacturers increasingly specialized in high-value tasks—R&D, design, branding—while importing labor-intensive assembly inputs from abroad (see Banga 2013). Passenger vehicles, electronics, and machinery all saw their foreign-input content climb, reflecting the growth of global networks of contract manufacturers and component suppliers (Kratz et al. 2024; National Science Board 2022).

Figure 4.1. Foreign-value-added in US Manufacturing Exports, % Manufacturing Exports

[INSERT FIGURE 4.1 HERE]

Notes: This plot shows the foreign-value-added in US manufacturing exports as a share of gross manufacturing exports from 1995 to 2020. This metric is calculated by dividing the value of imported inputs embodied in those exports by the total value of manufacturing exports in each year, then expressing the result as a percentage. The data is built for international comparison, so it uses a "standardized global system" to classify industries and is "globally balanced" to ensure all countries' trade adds up correctly.

Sources: OECD (2023b).

The Rise of an Intangible Economy

Figure 4.1 also indicates that, after roughly 2010, a declining FVA share was no longer driven by reduced fragmentation per se, but by a shift in the U.S. export basket toward knowledge-intensive goods and services whose value resides disproportionately in intangible assets.[63] While

[63] The post-2008 decline also partly reflects the aftershocks of the US Shale Oil boom in the form of rising domestic content in petroleum and coal related exports and changes in industry

supply-chain dispersion persisted, high-technology products with lower imported-input intensity than those associated with traditional heavy industry grew as a share of exports (see also National Science Board 2022; Borin and Mancini 2021). There is a good reason for this: between 2011 and 2020, it was U.S. firms that primarily (i) designed chips that were fabricated partly or entirely overseas, (ii) patented drugs in which the active ingredient was U.S.-sourced, and (iii) exported light machinery whose competitive edge lay in embedded software (Corrado et al. 2009; Haskel and Westlake 2017).[64]

The result was a supply-chain landscape that was still deeply globalized into the early 2020s—despite increasing tariffs and export controls (see Chapter 12)—but one whose disintegration was driven less by things like imported steel and plastics (Haskel and Westlake 2017; Corrado et al. 2009). Indeed, as globalization intensified, the US economy evolved from a manufacturing powerhouse into a service-oriented one dominated by industries built on intangible capital and human expertise. As manufacturing's share of GDP contracted, high-value services absorbed an ever-larger slice of America's export mix, with IP-rich know-how and

---

weights as energy prices and volumes moved; both mechanisms pull down the weighted-average FVA of manufacturing exports (OECD 2023a).

[64] Moreover, in the wake of rising offshore costs, longer lead times, and weak overseas demand, US firms reduced their reliance on foreign inputs in sectors like automotives and machinery (see MAPI Foundation 2017). At the same time, Chinese firms "up-streamed" assembly work for consumer electronics into higher-value manufacturing stages, while investments in advanced robotics and 3D printing enabled US factories to internalize more specialized, high-precision tasks (Kilpatrick et al. 2025).

digital intermediation driving a rising portion of US external earnings (White House CEA 2024). U.S. firms increasingly earned Euros, Pounds, Yen, Pesos, and Renminbi by selling or leasing IP-rich design, software, finance, and other business services as manufacturing supply chains became more internationalized. US firms cultivated a comparative advantage in all things "digital"; this category outpaced other service categories by the late 2000s (White House CEA 2024).[65]

Figure 4.2 conveys this phenomenon in hard numbers. In 1995, services of the business economy—those industries characterized by information and communications, financial intermediation, real-estate services, and professional and business support—accounted for just under 37 percent of total U.S. service exports. By 2020, that share had climbed above 55 percent, reflecting the nation's growing specialization in tasks founded on intangible capital: software platforms, data analytics, financial engineering, design services, and corporate consulting. That pattern zooms out from some informative examples: by 2020, the global payments processor Mastercard derived approximately 63 percent of its revenues from outside the US (Mastercard 2021), while Facebook (later renamed Meta Platforms) generated about 57 percent of its revenue from international markets (Facebook 2021).

And because Figure 4.2 ends in 2020, it misses some later milestones that underscore how cloud computing, digital marketplaces, and the globalization of professional expertise all

[65] To be sure, a non-negligible portion of these exports, especially in terms of IP, R&D, and business consulting, is boosted by corporate tax-avoidance strategies, in which US firms register profits in low-tax jurisdictions and then pay fees to their American parent—counted as services exports (see Setser 2020).

combined to elevate these business-economy services into the dominant engine of US export growth. By 2022, finance and insurance represented 27 percent of U.S.-owned multinational enterprises' service exports (BEA 2024), a reflection of how fintech platforms and cross-border digital transactions have surged. In 2024, total U.S. service exports exceeded $1.1 trillion, marking a substantial increase from prior years (BEA 2025). The U.S. trade surplus in services reached over $293 billion that year—approximately 26 percent of service exports—cementing the economy's transformation from one centered on physical goods to one powered by ideas, platforms, and relationships.[66] Indeed, in 2023, digitally deliverable services exports were $655.5 billion, and service export surpluses were largest in financial services and "other business services" (Stein et al. 2024).

Figure 4.2. High-Value Business Services Exports, % of Total Services.

[INSERT FIGURE 4.2 HERE]

Notes: the chart showing the share of high-value business-economy services (information and communication; financial intermediation; real estate; professional and business support) in total US service exports (which include construction).

Source: OECD (2023c).

[66] Despite its dominance in services exports, the U.S. also imports a significant volume of services. In 2024, U.S. services imports totaled $822.3 billion, reflecting demand for foreign travel, financial services, and telecommunications (Statista 2025). Key suppliers of U.S. service imports include the United Kingdom ($70.8 billion), Germany ($42 billion), and Canada ($40.6 billion) (BEA 2024).

The American economy transitioned to a more intangible capital-intensive economy throughout the same period, one centered on digital platforms, R&D-driven business models, and brand- or software-led value creation. Figure 4.3 reveals that, between 2002 and 2021, the Russell 3000's intangible-intensity—as measured by the share of property, plant and equipment invested in R&D, IP, brand equity, software and algorithms—rose dramatically, climbing from roughly 54 percent in the early 2000s to a peak above 95 percent by 2019.[67] After holding steady through the dot-com aftermath, intangibles began gaining traction in the mid-2000s, accelerating into the Great Recession; by 2008, more than three-quarters of the Russell 3000's capital stock was devoted to nonphysical assets.

From 2016 onward, the shift toward intangible capital picked up pace, reaching nearly 90 percent by 2018, and cresting close to 97 percent in 2019. The sharp drop back into the mid-80s in 2020 reflects both the COVID-19 shock and exceptional write-downs of physical plant in heavily impacted sectors, but the rebound toward 90 percent in 2021 shows intangibles' resilience—and their growing centrality—to corporate investment in America's largest publicly traded firms.

Figure 4.3. The Evolution of US Economy Towards Intangible Capital.

[INSERT FIGURE 4.3 HERE]

[67] That is not to deny that, during this era, R&D was also undertaken by public and quasi-public entities. These partnerships conducted joint research programs, shared computing resources, exploited data access agreements, participated in talent exchange programs, and undertook joint commercialization efforts (see Dean et al. 2023). It is to say, however, that this type of R&D was dwarfed by private sector spending (see Menaldo and Wittstock 2025).

Notes: The figure depicts intangibles as % of property, plant, and equipment for Russell 3000 companies. Namely, the percentage of intangible capital in the form of R&D, IP, brand equity (reputation/trust), computer software, and algorithms for the 3,000 largest publicly traded companies incorporated in the US between 2002 and 2021. To calculate these figures, I used Compustat data to identify the line-item balances for Property, Plant and Equipment and for Intangible Assets from each company's 10-K after consulting the Russell 3000 Constituent List from Russell Index LLC, which defines the 3,000 largest US-incorporated public firms for each year.

Sources: S&P Global Market Intelligence (2025) and FTSE Russell (2025).

The pattern revealed in Figure 4.3 is paralleled by the fact that, over this period, the US secured international cooperation over enhanced IP enforcement in venues such as the WTO (see Chapter 3), which helped American firms to generate profits from branding, IP, R&D, software, algorithms, and the commercialization of novel technologies.[68] US producers of ideas, platforms, processes, designs and services were able to specialize in creating these intellectual inputs, which were in turn either purchased or leased by foreign firms further down the supply chain, whether they were the manufacturers of inputs or downstream implementers. This underscores the fact that, save for a few exceptions, smaller and more nimble firms that are members of global innovation ecosystems occupy unique niches. Their comparative advantage tends to be tailored to only one segment of their products' supply chain: they specialize in either invention, or design, or testing, or packaging, or retail, rather than all simultaneously (see Barnett 2020).

[68] This paragraph draws on Menaldo and Wittstock (2025).

Figure 4.4, which graphs American IP exports between 1970 and 2015, both makes this clear and suggests that US firms largely occupy the upstream positions. From a baseline level indexed to one in 1970, net IPR receipts climbed only modestly through the late 1980s before rising sharply in the 1990s, roughly doubling by 2000. They then accelerated again in the early 2000s, driven by the enactment of TRIPS and related bilateral agreements, peaking at more than six times their 1970 level in the mid-2010s. This surge far outstripped GDP growth in the same period (shown by the dashed line), underscoring how much faster returns to IP expanded relative to the broader economy once global rules of the road for IP became more robust.

Figure 4.4. US Net IP Payment Receipts, % GDP.

[INSERT FIGURE 4.4 HERE]

Notes: Net IP receipts are obtained by subtracting incoming receipts from outgoing payments measured in 2010 dollars. GDP is expressed in 2010 dollars. Both are normalized to 1970, which equals 1. The vertical line, 2016, marks when the CDP era began to end after the election of Donald Trump to the US presidency.

Source: IMF Balance of Payments Statistics Yearbook and data files.

Big tech firms best embodied the hegemony of intangible capital in the American economy. Between 2020 and 2025, Alphabet, Amazon, Apple, Meta, and Microsoft each reinvested significant shares of annual revenues into R&D—cementing business models in which value creation increasingly flowed through software, data, and other intangible assets rather than physical capital (Alphabet Inc. 2020-2025; Apple Inc. 2020-2025; Meta Platforms Inc. 2020-2025; Microsoft Corp. 2020-2025). Chapter 8 will investigate the huge capital spending projects these companies embarked on to bankroll the AI explosion that occurred during the early 2020s.

THE ADVENT OF SUPERSTAR FIRM ECONOMICS

From the mid-1990s onward, the unbundling of every stage of production—design, component fabrication, software development, assembly, distribution, and after-sales services—into specialized, globally dispersed activities also enabled a handful of "superstar" firms to blossom. They concentrated their efforts and capital on creating, protecting, and monetizing intangible assets—patents, proprietary software platforms, brand equity, and data ecosystems—that yielded outsized returns. These returns were historically unprecedented.

This was a quintessentially American phenomenon. Gutiérrez and Philippon (2018) document that, between 1995 and 2015, the US average Herfindahl–Hirschman Index (HHI) at the broad sector level rose from roughly 500 to about 650, signaling a significant uptick in industry concentration. By contrast, ten large EU countries saw their HHIs fall from about 800 to 600 over the same period, and Valletti (2018) finds no comparable rise in European concentration.[69]

This concentration, while an imperfect proxy for the presence of superstar firms, is mirrored in rising profit shares and price-cost margins.[70] Shapiro (2018) shows that corporate

[69] To be sure, the US saw rising concentration from a low-concentration base, whereas Europe experienced de-consolidation from a high base.

[70] Industry-level concentration ratios such as HHIs are often invoked as proxies for firms' ability to price above marginal cost (Shapiro 1989), or to capture enduring competitive advantages (Porter 1980)—even though "industry" concentration aggregates across multiple, often heterogeneous product markets and geographic regions. Conversely, monopoly pricing occurs when a single firm faces the entire market demand curve and produces where marginal revenue equals marginal cost and Ricardian rents that allow some firms to produce for the market at

profits as a share of US GDP climbed from roughly 7.5 percent in 1985 to over 11 percent by 2016. Barkai (2017) argues that "excess profits"—returns to capital above what is needed to attract investment—surged during this era as risk-free rates plunged. Both point to increasing industry concentration as the reason why. Furman and Orszag (2018) document that the premium on private-capital returns versus safe assets swelled from about 200 basis points in 1985 to more than 800 by 2015, with those gains heavily skewed towards the largest, most productive firms. De Loecker et al. (2018) calculate that the average markup (price over marginal cost) among U.S. public firms rose from 1.21 in 1980 to 1.61 in 2016, driven almost entirely by top-quartile firms; Hall (2018) reports a similar jump in weighted average markups, from 1.12 to 1.39 between 1988 and 2015.

By the early 2020s, American big tech firms dramatically embodied the superstar dynamic. The combined market capitalization of Apple, Microsoft, Alphabet, Amazon, and Meta exploded, from under $300 billion in the early 1990s to over $6 trillion by the end of 2020.[71] By July 8, 2025, three American high technology firms had cleared the $3 trillion threshold: Nvidia

---

relatively lower marginal costs are possible in industries where many firms have varying levels of market shares. As a result, even a high sectoral HHI may mask fierce competition at the product-level (and vice versa), making it a noisy and imperfect indicator of the ability to price over marginal costs.

[71] As of December 31, 2020, Apple's market cap was $2.3 trillion (Macrotrends n.d.a), Microsoft's was $1.7 trillion (Macrotrends n.d.b), Amazon's was $1.6 trillion (Macrotrends n.d.c), Alphabet's was $1.2 trillion (Macrotrend n.d.d) and Meta's was $0.8 trillion (Meta Platforms Inc. 2021).

($3.9 trillion), Microsoft ($3.2 trillion), Apple ($3.1 trillion).[72] And if one considers Microsoft, Amazon, Meta, Nvidia, and Tesla together, in 2025 these firms together comprised roughly one-third of total S&P 500 market capitalization, a record concentration level (S&P Dow Jones Indices 2025). Unsurprisingly, these "magnificent seven" were highly profitable. For example, by mid-2025 Apple routinely posted operating margins north of 25 percent, versus a roughly 12 percent median for the broader index (Apple Inc. 2024a).[73]

The Underlying Logic Behind Superstar Firms

What explains the rise of America's superstar firms earning outsized profit margins? [74] Several answers, all of which align with the bigger story of the US economy I've told so far in this chapter, have been put forward. They include superior intangible assets and supply chains, barriers to entry, and inimitable productivity advantages. Crouzet and Eberly (2019) point to the growing importance of intangible capital that thrives under scalable, contract-driven, disintegrated supply chains. Autor et al. (2020) claim superior products and productivity allowed the "winners" to capture ever-larger market shares. Galetovic (2021) and Gochberg and Menaldo (2022) posit sector leaders earn outsized Ricardian Rents due to superior technology or knowhow or both: they exploit the wedge exists between their higher productivity/lower marginal costs and the market price.

---

[72] See Macrotrends n.d.e; Macrotrends n.d.f; Macrotrends n.d.g). Meanwhile, Alphabet was worth $2.116 trillion (see Macrotrends n.d.h).

[73] The same was true of Alphabet (see Alphabet, Inc. 2025).

[74] Cross-country work by Calligaris et al. (2018) and Diez et al. (2019) show only modest overall increases in markups over the same period, therefore pointing to US exceptionalism.

Allow me to unpack the last explanation. Consider that in a competitive industry with free entry and exit, and where no firm has market power, the costs of production are (in the short term at least) increasing as a function of the quantity of the product in question. Critically, this is also true for the long-run supply curve in an increasing cost industry when productivity differentials between relatively more productive firms and the least efficient one that are never fully eliminated. This is because the equilibrium price in a competitive market characterized by Ricardian Rents is determined by the least efficient producer, which produces the marginal quantity where the supply curve and demand curve intersect and is basically indifferent between remaining in the market and exiting it.

In other words, it is that firm's unique marginal cost curve, and specifically where it intersects with the market demand curve (at the minimum of its average cost, where the marginal cost equals the average cost) that is going to dictate the price for all market participants. Because inframarginal (more productive) firms sell at that same market price, they earn Ricardian rents equal to the wedge between price and their own costs—even when they compete in an arena defined by competitive prices, where consumer surplus is maximized, and deadweight losses are non-existent (see Schmalensee 1982 and Peteraf 1993).

The distinction between monopoly rents and Ricardian Rents has a profound implication that I will revisit in Chapter 7, when this book explores how antitrust law and enforcement has addressed issues around creative destruction in general and digital platforms in particular.[75] In

[75] However, as I will demonstrate in Chapter 11, these two distinct micro-foundations produce the exact same macroeconomic result in the model I introduce there: a "decoupling" of average wages from productivity. Indeed, in that chapter, I will subsume both phenomena under a single

the latter case, a firm's markup is capped by competition rather than demand elasticity. This distinction can be formalized using the notation I will develop fully in Chapter 11. A traditional monopolist prices according to the Lerner Rule (Lerner 1938), which states that a firm's profit margin depends on how price-sensitive its customers are (elasticity, or ε). Formally, the Lerner Index (L) is defined as:

$$L = \frac{(P - MC)}{P} = \frac{1}{\varepsilon} \quad (4.1)$$

Where:

- P is the price
- MC is the marginal cost
- ε is the elasticity of demand.

If I rearrange this formula to solve for the monopolist's markup ($\mu^{\text{Monopoly}} = \frac{P}{MC}$), a monopolist sets price as a multiple of cost defined purely by demand. That is, they raise prices until a critical mass of consumers threatens to stop buying:

$$\mu^{\text{Monopoly}} = \frac{\varepsilon}{(\varepsilon - 1)} \quad (4.2)$$

---

parameter. Readers should therefore note that when I formally model "market power" later in the book, I am not necessarily accusing superstar firms of restricting output to gouge consumers. Rather, I am capturing the structural reality that when Ricardian Rents are high, the math of the economy changes in a similar way as to when monopolists price as $Eq.4.2$ implies: the economywide profit share of GDP rises, and the share flowing to labor and capital necessarily falls, regardless of whether that outcome is driven by monopoly pricing or superior efficiency.

However, the superstar firms of this era were often engaged in Limit Pricing (see Bain 1956; Milgrom and Roberts 1982). They could not raise prices to the theoretical monopoly level because they were constrained by a "competitive fringe"—rivals ready to enter the market if prices rose too high. In this scenario, the market price (P) is anchored by the marginal cost of the inefficient fringe firm (MC_{fringe}) and the superstar's markup (μ^{star}) is determined strictly by its "Relative Efficiency":

$$\mu^{\text{star}} = \frac{P}{MC_{Star}} \approx \frac{MC_{fringe}}{MC_{Star}} \tag{4.3}$$

Therefore, $Eq.\,4.3$ reveals that, unlike the monopolist's markup, $\mu^{\text{Monopoly}} = \frac{P}{MC}$, the superstar's profit margin, $\frac{P}{MC_{star}}$, is not a tax on consumers; it is a measure of the "Efficiency Gap" between the leader and the laggard.[76] I will return to this formalization in Chapter 11, when

[76] Strictly speaking, Ricardian rents arise in competitive markets where firms are price takers and the market price is set by the marginal cost of the least efficient producer. However, in the case of "Superstar Firms" with dominant market shares, I apply the logic of Contestable Markets (see Baumol et al. 1982). In this framework, the Superstar acts as a price maker but acts as if it faces competition: it engages in Limit Pricing by setting its price just below the marginal cost of the inefficient "competitive fringe" (MC_{fringe}) to deter entry. Consequently, while the firm may superficially resemble a monopolist in terms of scale, its excess profits are technically Ricardian Rents—derived from its superior efficiency relative to the fringe ($MC_{fringe} - MC_{Star}$)—rather than Monopoly Rents derived from restricting output to raise prices above the competitive ceiling.

revisiting markups and their role in determining the distribution of income in the AI economy and the possibility that it may drive a wedge between workers' wages and labor productivity.[77]

Perhaps the easiest way to grasp this is with a simple analogy. Imagine several farmers with different qualities of land all selling wheat to the same town. One farmer has incredibly fertile, well-irrigated land and can produce a bushel of wheat for just $3. Other farmers have less productive land, with the last and least efficient farmer needing to spend $5 to produce a bushel. To meet the town's total demand, the market needs the output of all the farmers, including the least efficient one. Therefore, the market price for "all" wheat will settle at $5 per bushel. If the price were any lower, the least efficient farmer would stop producing, and there wouldn't be enough wheat to go around.

In this scenario, the least efficient farmer makes zero economic profit—they sell at their marginal cost. However, the most efficient farmer, whose costs are only $3, earns a handsome profit of $2 on every bushel they sell. This $2 profit is their Ricardian Rent. It's a reward not for monopoly power, but for their superior, more productive asset—the fertile land.

The important point, therefore, is that these rents are not a byproduct of scarcity or uniqueness in an absolute sense, only a relative one. This is what allows several efficient producers can earn Ricardian Rents in a fully competitive market if they enjoy differential productivity vis-à-vis other producers. The deeper reason for this is that access to high quality inputs, technology, talented management, and human capital are not uniformly distributed. Moreover, resources that furnish more efficient producers with a competitive edge cannot be

[77] Following Barkai (2020), I demonstrate there that high markups may siphon value away from both Labor and Capital.

easily copied or learned by other firms even when technology transfer between firms is relatively high.

Sources of Ricardian Rents

Amit and Schoemaker (1993) argue that certain strategic assets provide firms with sustainable competitive advantages. The asset must demonstrate scarcity, meaning it exists in short supply relative to market demand. It should have low tradability or limited transferability, preventing it from being readily bought or sold on factor markets. The asset must also exhibit inimitability, where rivals cannot copy it at reasonable cost due to factors such as causal ambiguity, complexity, or history-dependence. Limited substitutability is another crucial element, requiring that few close functional substitutes exist for the asset. The firm must also be able to maintain appropriability, meaning it can capture the returns the asset generates through mechanisms like patents or trade secrets. Durability plays an important role as well, with the asset depreciating slowly such that its economic life exceeds rivals' imitation or learning speed. And the asset should demonstrate complementarity, where its value increases when combined with the firm's other assets, including co-specialized assets. Finally, there must be overlap with strategic industry factors, ensuring the asset aligns with the factors that drive profitability in the industry at a particular point in time, making the asset relevant to competitive dynamics.

This take aligns with Bessen (2020), who underscores that proprietary IT investments—software and process innovations used internally—can confer persistent advantages that ripple across industries. Corroborating this thesis, Bessen and Calligaris (2018) show that firms' share of "IT intensity" (proprietary IT as a fraction of net investment) jumped from 7 percent in 1985 to 24 percent in 2016, roughly matching R&D expenditures and cementing the gap between frontrunners and laggards. Firms that were early adopters of computers, software, intra networks,

the internet, the precursors to cloud computing, and an array of digital technologies were able to gain palpable advantages over their rivals.

Examples during the 1990s and early 2000s include Walmart in the Department Stores category, Jet Blue in the Airline sector, and Exxon-Mobile in petroleum and natural gas. Walmart forged several innovations in supply chain management and warehousing, including the widespread use of barcodes, computer networks, and data storage, which trickled out to other firms in its industry, such as Target. These pack-leading retailers exploited their IT advantage to forecast sales with great accuracy, slash their inventories—and thus cut down on waste and reduce costs—adopt uniform standards for labelling products, and create interoperable production and sales information that could be automatically shared in a standardized manner across their global supply chains and between their retail outlets (see Bonacich and Hardie 2006, p. 108).

Later, it was Big Tech firms that exploited big data and groundbreaking algorithms to obtain both direct and indirect network effects that gave them outsized advantages. This was exemplified by Google's dominance in digital advertising linked to internet search, Amazon's pioneering logistics systems and cloud computing offerings, and Netflix's recommendation engines and viewer feedback-based content offerings. Across these cases, proprietary data and custom-built algorithms created a self-reinforcing feedback loop, where a better, more personalized user experience attracted more users, which in turn generated more data to further refine the algorithms and widen their competitive moat.

Returning to the idea that superstar firms earn Ricardian Rents, we can consider these high-tech companies' unique and hard-to-copy intangible assets as their "fertile land." Just as the best land is scarce, strategic assets such as proprietary software (e.g., Google's search algorithm),

unique datasets, superior R&D talent, or a trusted brand, are not uniformly distributed and cannot be easily replicated by competitors. Data assets and the models trained on them are scarce, inimitable, and durable. Moreover, cloud computing architectures, specialized chips, and software are complementary and often co-specialized—yielding rents without requiring monopoly pricing.[78]

Finally, I should emphasize the inherent fragility of the "superstar phenomenon," in which the source of rents also ironically sows the seeds of its own undoing. While in Chapter 11 I will formalize how technological diffusion threatens these firms' rarified perch, later in this chapter I will demonstrate empirically that high-tech superstars acted as "hubs" that diffused technology to the rest of their respective industries. This creates a paradox: as diffusion occurs, the fringe adopts the leader's techniques, lowering MC_{fringe} and compressing the leader's markup defined in Equation 4.3. Thus, to sustain these "Ricardian Rents," the superstar is forced onto a treadmill of constant innovation—they must lower their own costs (MC_{Star}) faster than diffusion lowers their rivals' costs. Such are the vagaries of creative destruction!

Why European Firms Did Not Achieve Superstar Status

Companies such as Apple, Google, and Amazon were able to supercharge innovation and achieve unprecedented scale, allowing them to extract persistent profits by monetizing data and

[78] This strongly contrasts with the view that digital platforms' indirect and direct network effects confer unassailable market power onto them by increasing switching costs and representing steep barriers to entry (see Katz and Shapiro 1994; Rochet and Tirole 2003). The high degree of multihoming—consumers' switching back and forth between multiple platforms—is a prima facie challenge to that traditional view (see Armstrong 2006; Evans and Schmalensee 2016).

building engaging digital ecosystems built on R&D and acquisitions. Conversely, Europe's policies and regulations favored more traditional manufacturing populated by vertically integrated firms and a more risk-averse business culture that forestalled a transition to an economy centered on software, data, and algorithms (see Bradford 2023). The EU and European countries' regulatory environment—through data protection, digital market acts, and industrial policy—resulted in fewer globally scaled digital platforms or firms specializing in intangible capital and design. This is a topic I will return to in Chapter 12.

Meanwhile, Japanese Superstars like Toyota and Sony achieved "Global Frontier" status through fierce competition and operational brilliance, their efficiency never diffused to the domestic service sector (Katz 1998; McKinsey Global Institute 2000; Katz 2025). Protected by regulations that stifled competition, the median Japanese firm failed to adopt the digital tools of the leaders. The result was a "Great Divergence": while the Superstars raced ahead, the rest of the economy stagnated, creating a "High-Technology/Low-Productivity Trap' where the benefits of innovation remained locked inside a few corporate giants.

In the next chapter, I will revisit this example in light of empirical evidence from the U.S. high-tech sectors that tells a different story. The chapter maps an equally inegalitarian yet high diffusion equilibrium conceptualized as an "innovation commons."

CONCLUSION

This chapter charted a profound economic transformation, tracing the arc from the postwar era of renewed globalization to the rise of a U.S. economy built not on factory floors but on intangible capital. The international economic architecture championed by the United States—defined by the dollar's hegemony, open capital flows, and expanding trade networks—reshaped global

commerce. It catalyzed the vertical disintegration of supply chains, allowing American firms to specialize in high-value, IP-intensive activities like design, R&D, and software development.

This structural shift, in turn, paved the way for the quintessentially American phenomenon of the “superstar firm.” As I have argued, these firms did not rise merely through anticompetitive consolidation, but by leveraging scalable intangible assets and superior "business acumen" to capture outsized market shares. By utilizing disintegrated supply chains to focus on their core competencies, they achieved productivity levels that far outstripped their peers, earning "Ricardian Rents" based on their relative efficiency rather than monopoly tax.

This economic pivot was supported by the antitrust shift I explore in Chapter 7, where US courts moved away from a reflexive hostility toward large firms to recognize that market dominance, when earned through superior business acumen, is a pro-competitive outcome.

However, as the contrasting cases of Europe and Japan reveal, this ecosystem is fragile. Europe’s risk aversion and regulatory hurdles stifled the emergence of such firms entirely, while Japan’s protectionism allowed its superstars to rise but prevented their innovations from lifting the broader economy. These counterfactuals underscore a critical lesson: the "superstar" economy only delivers broad prosperity if the "frontier" of innovation eventually diffuses to the "fringe."

CHAPTER 5: HIGH TECH SUPERSTAR, INDUSTRY TECH LEADERS

In the popular imagination, the dominant firm is often cast as a hoarder—a dragon guarding a fortress of intellectual property designed to starve competitors and stifle the market. Yet, the history of the Third Industrial Revolution reveals a reality that defies this caricature. In the US, the era that birthed the personal computer and the mobile internet was not defined by walled gardens, but by a vibrant, invisible architecture of shared knowledge. Technological leaders cemented their dominance not by hiding their inventions, but by broadcasting them, turning their proprietary breakthroughs into the industry's common language.

Conversely, during this same period Japan's "superstars" became isolated islands rather than central hubs. While giants like Toyota and Sony achieved "global frontier" status through continued innovation and fierce efficiency, their exported technologies failed to diffuse to the domestic sector, which was shielded from competitive pressures. This Balkanization took the form of a stagnant "dual economy" (see Katz 1998; McKinsey Global Institute 2000; Katz 2025).

To understand how US superstar firms in high-tech sectors became their industries' innovation leaders during the Third Industrial Revolution, consider a parallel example from biotechnology. The Cohen-Boyer patent for recombinant DNA, the foundational technology of the entire biotechnology industry, was filed in 1974 by Stanford and the University of California. The universities' strategy of offering broad, non-exclusive licenses spurred massive horizontal technology transfer, as hundreds of startups and established pharmaceutical firms licensed the patent to build the modern biotechnology industry. This phenomenon drove vertical technology transfer as well, creating a new market for specialized suppliers of enzymes, lab equipment, and other tools needed to perform gene-splicing methods.

This chapter shows that innovation in high-tech industries during this era took on a similar networked character. However, unlike biotechnology, where universities often took center stage, in this innovation commons firms collaborated and competed to create the foundational technologies of the digital world. In the open, clique-free structure of the US semiconductor network, "superstars" were highly connected conduits for broad-based diffusion. For example, Intel was responsible for introducing and standardizing technologies like the CPUs that power laptop computers. It and other firms like it led collective efforts to produce the critical precursors to the modern smartphone, even though there was often no formal system that orchestrated technological creation and technology transfer from the top down.

Inequality Does not Mean Hoarding

Consider the distribution of city sizes and what it might tell us about innovation commons that are highly polarized between leaders and followers. Almost every one of the world's countries boast one or two mega-cities, a handful of large cities, and then a vast number of small towns and villages. The defining feature of this type of distribution is a "fat tail," which means that extreme events—like a city the size of New York—are not impossible outliers but are an expected part of the distribution.

Ahead, I introduce several datasets that map the technological relationships between firms and deploy network data analysis to explore how patenting networks spread technological advancements in the semiconductor sector and the other high-tech sectors while embedded within very similar "Pareto" distributions.[79] Leading firms act as important hubs within dense

[79] While data hosting/processing and computer design services were emerging during this period, I exclude them from this analysis. In the pre-2007 era, these sectors were primarily business-to-

networks of technological learning, holding several influential patents with numerous forward citations—reinforcing their status as technological leaders.

A prime example of this dynamic can be found in the rise of Texas Instruments (TI) as a hub for digital signal processing technology. In the 1990s and 2000s, TI developed a portfolio of foundational patents for digital signal processors (DSPs), specialized microchips essential for handling real-world signals like audio in the burgeoning digital mobile phone and consumer electronics markets. This spurred horizontal technology transfer, as rival semiconductor firms had to either license TI's technology or invest heavily to design around its foundational patents to compete. Simultaneously, it drove vertical technology transfer down the supply chain, as major device manufacturers like Nokia and Motorola depended on TI's DSPs and licensed its patents to build the core functionalities of their mobile phones.

Moving beyond this example, the chapter combines descriptive evidence and formal statistical analysis to tell this story systematically. I operationalize technological leadership using firms' patent forward citations. The data reveal that high-tech industries conform to a Pareto

---

business service providers, like the systems integrator Electronic Data Systems (EDS), which specialized in designing, installing, and managing large-scale, custom IT systems for corporate clients, and remained peripheral to the core consumer-facing supply chain for PCs and early mobile devices. Eventually, data hosting evolved into the cloud computing infrastructure that underpins the entire AI ecosystem by storing and processing the massive datasets required for machine learning, while computer design services became the architects that build and deploy the Internet of Things (IoT) and integrate AI solutions into modern enterprise operations. I will revisit these contributions to the Fourth Industrial Revolution again in Chapter 9.

distribution in which the top 20 percent of firms account for nearly 98 percent of all forward citations—an unmistakable “winner-take-most” pattern that distinguishes the sector from the rest of the economy. Firms that dominate citation networks consistently exhibit TFP levels roughly 70 percent higher than their peers.

Network analysis demonstrates that innovation flows through a small number of highly connected semiconductor hubs—Texas Instruments, Intel, and Qualcomm—that diffused foundational technologies to a wider periphery of adopters. While these high-tech leaders were remarkably effective at propagating their own general-purpose technologies, they were less likely than non-high-tech firms to act as “curators” of external innovations. This insularity was not a weakness but a defining feature of their dominance: rather than importing best practices from other sectors, *they became the best practice*, setting the technological standards that underpinned the transition to the digital economy.

However, I also argue and show that inegalitarianism did not preclude a "win-win dynamic" benefiting both leaders and followers. In fast-moving and novel technology spaces, both types of firms together sought to create new markets by jointly shifting out the demand curves for embryonic digital electronic devices. Their biggest challenge was to persuade consumers to acquire and learn how to use operating systems and applications and it was only when supply chain partners and even bitter rivals worked together that they could hope to create an interoperable ecosystem of devices, applications, and complementary markets. The industry's success hinged on coordinated, positive-sum effort across different layers of its technology stack.

Take early PCs, for example: their complexity limited their appeal until a new generation of machines powered by faster processors came bundled with "Windows 95" and web browsers. This combination transformed the PC from a difficult-to-use office machine into an easy-to-use

gateway to the internet, e-mail, and multimedia—a massive surge in value that shifted the entire demand curve. Intel's relentless push for more powerful microprocessors set the pace; each new generation of chips enabled more demanding software, creating the foundation for Microsoft's Windows 95 and the first web browsers. PC manufacturers like Compaq and Dell competed to bundle this hardware and software into affordable machines, creating a much larger pie for the entire industry.

USING PATENT DATA TO MAP THE INNOVATION COMMONS

I use patent data to document these technology creation and diffusion dynamics more precisely. When a new patent application cites the "prior art" it builds upon, these references are called backward citations. The more telling metric for our purposes, however, is a patent's forward citations. Just as a paper's true impact is measured by how many future scholars cite it, a patent's influence is measured by the number of subsequent inventions that reference it. This makes tracking forward citations instrumental in identifying technological leaders and mapping the flow of ideas (Henderson et al. 1998; Brantle and Fallah 2007; Hall et al. 2001).

Importantly, there are sincere rather than merely strategic reasons why inventors cite previous patents. Not only does an applicant face a legal duty to disclose knowledge of prior art, but a professional patent examiner conducts a thorough search and helps the applicant identify the most relevant patents to cite. For these reasons, forward citations should be reliable indicators of an invention's importance and impact, and the process of acknowledging prior art represents a powerful proxy for actual technological links between inventions (Jaffe et al. 1993; Caballero and Jaffe 1993).

To map patent citation networks for the technologies instrumental in driving the Third Industrial Revolution, I leverage the NBER "Utility Patent and Patent Citation Data File" (Hall et

al. 2001), which identifies patents granted by the USPTO between 1976 and 2006—critically, the eve of Apple's introduction of the iPhone. I aggregate individual patents to publicly traded firms using the COMPUSTAT database, focusing on firms existing between 2000 and 2006. These efforts yielded a cross-sectional dataset of 25,480 firms holding over 3 million patents that received more than 28 million forward citations.[80]

I took several steps to ensure these findings were not merely artifacts of firm longevity or survival bias. The 2000–2006 firm window anchors the analysis in the mature pre-iPhone ecosystem while capturing the long innovation run-up via patents granted since 1976. Because patent counts and citations mechanically accumulate with time—and because citation practices vary across technological vintages—I normalize key measures by firm age, with a common start-year floor of 1950 for older incumbents. Furthermore, the analysis excludes firms that exited the dataset between 2001 and 2007 to ensure that the identified "hubs" were not merely legacy holdovers, but active participants in the Third Industrial Revolution's peak. Finally, I applied a "singleton restriction," dropping sectors with only one firm; since, by definition, a firm cannot lead or follow itself nor, by extension, exchange knowledge with itself.[81]

---

[80] Menaldo 2026 (Section 2) contains a longer version of this chapter and explains the technical details in greater depth, performs several additional statistical exercises, and provides richer case studies of the firms discussed here.

[81] Other things to note. The analysis focuses exclusively on utility patents, which protect how an invention works, as opposed to design patents covering ornamental appearance. Where patents have multiple assignees (about 2 percent of the dataset), I assign the patent to each listed firm. Finally, I identify high-tech sectors versus non-high-tech sectors. To qualify as "high-tech," a

EVIDENCE OF SUPERSTAR DYNAMICS

While the relationship between technological leadership and productivity is present throughout the history of high-tech (see previous chapter), multivariate regression models provide compelling evidence. After holding constant sector-wide characteristics such as R&D spillovers and average patent density, there is a strong, statistically significant correlation between technological leadership measured as patent forward citations and TFP: 67% ($p < .001$); $n = 5{,}060$.[82] Even more telling is what happens after I apply sector fixed-effects. After "expunging" all sector-level omitted variables—essentially comparing a leader to its immediate peers within the same sub-industry—the results remain materially identical.

Simpler descriptive statistics immediately signal that this productivity advantage is concentrated in a small handful of pacesetters that became massively influential hubs, while the

---

firm must be classified by the NAICS as either operating in wireless telecommunications, semiconductors, computer manufacturing, software publishing, data processing and hosting, computer system design services (Network Systems Integration Design Services), other computer services, and custom computer design services (Computer Program or Software Development).

[82] TFP is from Bahar (2018), estimated as the residual from a Cobb-Douglass production function using data on operating revenue, tangible fixed assets, number of employees, and the cost of materials. I calculate firm-level averages between 2007 and 2015 for all public companies in the COMPUSTAT dataset, allowing for a clear assessment of how the technological leadership established in the preceding decades (1976–2006) translated into long-term productivity advantages. See Menaldo (2026), Section 2 for how I cluster the standard errors.

vast majority of firms remained on the periphery. Technically, a quantity, x, is governed by a power law if it is drawn from a probability distribution where $p(x) \propto x^{-\alpha}$, where the "tail" of the distribution becomes heavier as $\boldsymbol{\alpha}$ gets smaller. If we take logs of both rank order and frequency, we should observe a straight line on a doubly logarithmic plot. If that's the case, in all likelihood the patent citation distribution for high-tech firms follows a Power Law distribution—a "rich-get-richer" phenomenon whose most famous statistical feature is a characteristic "fat tail," meaning massive outliers are not just possible but expected.

[FIGURE 5.1 HERE]

Figure 5.1: Power Law Fitted to Patent Forward Citation Data, High Tech Firms.

Notes: both the y and x axes are logged. I add 1 to the number of patents weighted by firms because several firms are observed as having zero patents. On the y-axis, I list the lowest value (0+1), 25th percentile, median, 75th percentile, 90th percentile, 95th percentile, 99th percentile, and highest value. On the x-axis, I list the 25th percentile, median, and highest value.

The log-log plot in Figure 5.1 indeed reveals the expected linear relationship, particularly for firms in the right tail of the distribution: those with higher forward citation counts. This "straight-line" diagnostic is the empirical fingerprint of a Power Law, suggesting that "Innovation Hegemons" are not random outliers from a normal distribution, but the natural focal points of a system governed by a strict hierarchy separating the leaders of the pack from the followers.

Several tests confirm this expectation. The overall inequality is staggering, associated with a Gini coefficient of .96, which is consistent with the finding that the top 20% of high-tech firms command 98.8% of all forward citations. Moreover, this extreme concentration is indeed driven by a Power Law dynamic: the data obeys a Pareto distribution above the 75th percentile (a

threshold of 218 citations); I fail to reject the null hypothesis that it is not distributed in this manner (p-value = .40).

This headline finding is robust. First, to how I measure technological prowess: when I use average citations per patent, the results are identical. Importantly, this is not a universal feature of the American economy: exploiting the goodness-of-fit technique pioneered by Urzúa (2020), while the patent distribution for non-high-tech firms fails to conform to a power law at any threshold, I fail to reject the hypothesis that the high-tech firms obey a power law distribution, including at very high thresholds (see Menaldo 2026: Section 2).

THE SEMICONDUCTOR NETWORK: MAPPING TECHNOLOGY DIFFUSION

To identify the distinct "innovation leaders" that developed and diffused inventions central to their industry, I constructed network datasets observing citation ties between high-tech firms existing between 2000 and 2006. The first is a "horizontal network" observing patent citations among 127 semiconductor firms—a directed network where a tie is recorded if one firm cites any of another firm's patents at least once. The second is a “vertical network” that represents the PC/internet supply chain, with patent citations linking players in different segments of the stack.

In the citation networks that follow, each node represents a firm and each directed edge represents a citation relationship: Firm A's patents cite at least one patent assigned to Firm B. Hub firms are those with high in-degree (many firms cite them), while bridge firms score high on betweenness centrality (they sit on many shortest paths connecting otherwise distant firms). Clustering captures whether firms tend to cite the same sources; the degree distribution reveals whether influence is broadly shared or concentrated in a few hubs.

Why networks? Patent citations leave a paper trail of technical dependence: firms cite prior inventions when they build upon them, solve around them, or adapt them into new

products. Mapping these citation ties as a network allows us to test a specific theory of technological diffusion—whether foundational ideas spread broadly through an ecosystem, and whether that spread is mediated by a few key hubs rather than confined within isolated silos. Figure 5.2 provides a global visualization of this horizontal network. The network's most striking feature is its heterogeneous connectivity pattern, a core-periphery structure with dominant hubs at the center. Texas Instruments is cited by 99 firms (78 percent of the network); Intel by 89. At the periphery, 25 firms receive no citations at all, while 47 receive five or fewer. Between the hubs and the far periphery sits a middle tier of moderately connected firms that help relay and recombine core innovations as they diffuse through the ecosystem.

[FIGURE 5.2 HERE]

Figure 5.2: Visualizing the Semiconductor Network.

Notes: This network visualization displays the citation relationships among 127 semiconductor firms through 2,347 directed ties (links) based on patent citations from 1976 to 2006. Node size is proportional to the number of times a firm is cited by other semiconductor firms in the network (in-degree centrality). The largest nodes represent the most influential technology hubs: Texas Instruments (cited by 99 firms), Intel (cited by 89 firms), and Qualcomm (cited by 39 firms). The network exhibits a core-periphery structure with a density of 0.15, meaning that only about 15 percent of all possible citation ties exist. Twenty-five firms receive no citations, while 47 firms receive five or fewer citations, appearing as the smallest nodes on the periphery. A tie is recorded as a binary relationship: a link exists if one firm cites any of another firm's patents at least once during the analysis period, irrespective of the total number of citations.

Sources: NBER Utility Patent and Patent Citation Data File (Hall et al. 2001); S&P Global Market Intelligence COMPUSTAT database (2025).

Even though the network is sparse, firms are only a few steps apart, which makes diffusion structurally easy: a foundational idea originating in one part of the semiconductor ecosystem can reach most other firms through a small number of intermediate links (see Menaldo 2026, Section 2 for metric definitions and values).

High clustering coefficients combined with low transitivity reveal a telling structural pattern. The average clustering coefficient of .61 indicates that firms tend to cluster around common central sources of innovation. Local clustering is high, but global closure is not. Yet the network is not abnormally "cliquish"—its transitivity score aligns with what chance alone would produce. This means knowledge flows through central hubs rather than getting trapped in isolated, self-referential subgroups. In social network terms, this is less like a group of friends who are all friends with each other (a closed clique), and more like a group of students who all seek to befriend a network influencer. Notably, this pattern is consistent with hubs that broadcast foundational technologies widely rather than anchoring closed, mutually-referential circles.

The network's node degree distribution confirms the Power Law structure identified in the cross-sectional data. Figure 5.3 plots the probability degree distribution on a log-log scale; the predicted line becomes increasingly linear in the right tail, and formal testing using the Urzúa (2020) technique confirms that the data conforms to a power law above the 75th percentile of the distribution (p-value = .11).

[FIGURE 5.3 HERE]

Figure 5.3: Power Law Fitted to Probability Degree Distribution

Notes: The predicted line for log(number of ties + 1) is estimated using a running line smoother. I estimate each smooth by employing a backfitting algorithm and a running-line smoother for the log(probability of the observed number of ties occurring). The smoother is a linear function for

each observation of that variable and 95 percent confidence intervals are calculated from standard errors for each smooth.

The Big Three: Texas Instruments, Intel, and Qualcomm

Three firms emerge as the semiconductor network's clear leaders. They played instrumental roles across different segments during the analysis period: Texas Instruments for analog/embedded processing; Intel for microprocessors; and Qualcomm for wireless technologies. Despite different specializations, they shared a similar playbook: aggressive risk-taking, massive R&D investments, and pioneering new models of industrial organization.

In diffusion terms, these hubs are not just "big firms." They function as suppliers of general-purpose platform technologies that become shared building blocks—technical languages that thousands of downstream and peer innovations must speak to remain compatible, competitive, and manufacturable at scale. This is precisely why their patents become common reference points across otherwise specialized and diverse segments of the semiconductor industry.

Texas Instruments pioneered Digital Signal Processors (DSPs), specialized chips that became the engine of the digital media and communications revolution. US Patent No. 4,577,282, granted in 1986 for the architecture of its landmark TMS320 family of processors (Caudel and Magar 1986), was essentially a specialized computer on a single chip, optimized for processing real-world analog signals. The DSP became the heart of early digital cell phones, high-speed modems, and hard disk drive controllers. Nokia relied on Texas Instruments' DSP cores in its digital mobile phones throughout the 1990s, using them to convert voice signals into digital form for 2G networks. The patent was cited 66 times between 1976 and 2006, placing it comfortably in the top few percent of all patents.

Texas Instruments ranks first in network influence, with over 2,800 different patents cited more than 40,000 times by semiconductor peers alone. Its betweenness centrality score exceeds 1,800 (versus a median of 1.5), indicating it is not just a source of innovation but the chief conduit through which technology transfers between other firms. Its network-leading eigenvector centrality score of 0.20 confirms that it is cited by other important, central players.

Intel secured landmark patents critical for PC performance, including US Patent No. 5,584,001 on branch prediction in superscalar, out-of-order cores (Hoyt et al. 1996). A "superscalar, out-of-order" processor executes multiple instructions per clock cycle and reorders them to keep components busy. The challenge is "branch" decisions in programs that create uncertainty about which instructions to execute next. Intel's innovation gave processors a sophisticated system for making educated guesses about branch outcomes, using a "branch-target buffer" to keep detailed history of past outcomes and improve accuracy over time. By drastically reducing pipeline stalls, this innovation produced the dramatic leap in real-world speed essential for graphics-heavy software like Windows 95.

Intel is the network's second most influential hub, cited by 89 firms with over 102,000 total forward citations. Its betweenness centrality of 1,556 confirms it acts as a critical bridge for technology transfer, with over 1,500 of the network's shortest communication paths flowing through it.

Qualcomm, one of the first "fabless" firms to focus purely on designing chips fabricated in Asian foundries, designed critical modem chips for 2G and 3G mobile phones. Its most cited patent, US Patent No. 5,103,459 (Gilhousen et al. 1992), describes Code Division Multiple Access (CDMA) technology—a revolutionary method for allowing multiple mobile phones to share the same radio spectrum simultaneously without interference. Unlike older methods

dividing airwaves by frequency or time slots, CDMA gives each user a unique digital code. This "spread spectrum" technique was more secure and offered greater capacity, making it foundational for mass-market wireless communication. The patent amassed 917 citations, placing it in the 99.99th percentile.

Qualcomm's position is more specialized but strategically critical. With only 39 semiconductor firms citing it, it ranks just 19th in direct network ties. However, it exhibits the highest average patent generality score (.65) among top-cited firms, indicating its technologies have broad applicability across domains. Statistical analysis confirms that firms citing Qualcomm are the most dissimilar to it in patent generality—specialized firms specifically seek out Qualcomm for access to foundational, broadly applicable wireless technologies.

Technology Flows from Hubs to Periphery

A key question is whether peripheral firms seek access to foundational technologies from network hubs. I measure patent "generality"—scores ranging from 0 to 1, with values approaching 1 indicating the patent is cited broadly across many technological domains (suggesting general-purpose technology characteristics). A high generality score for forward citations means a patent influenced follow-up innovations across a variety of fields.

Statistical analysis confirms that "opposites attract." For Texas Instruments, there is a significant negative correlation (-0.20) between a firm's similarity to Texas Instruments in patent generality and its likelihood of citing it. Firms with specialized, low-generality patents were most likely to cite the foundational, high-generality patents of Texas Instruments. The same pattern holds for Intel (-0.20) and is even stronger for Qualcomm (-0.24). Conversely, there is no significant relationship for ties between other firms. This provides strong evidence that peripheral "spoke" firms specifically seek out "hub" firms for general-purpose technology.

This pattern is difficult to reconcile with a “siloed innovation” view in which specialized firms primarily build on other specialists (a homophily story). Instead, it fits a diffusion mechanism in which peripheral innovators deliberately reach into the core to acquire broadly enabling, general-purpose technologies that can be adapted into many specialized applications.

If a common set of “bedrock” technologies is organizing the ecosystem, we should see many of the same firms citing the same upstream leaders. To test that implication, I examine whether the leaders share a common citation “audience”—that is, whether the same semiconductor firms repeatedly cite each of them. The leaders do show substantial overlap. For example, the citation networks of Texas Instruments and Intel overlap substantially, with a correlation of 0.65, indicating they are frequently cited by the same set of semiconductor firms (see Menaldo 2026, Section 2 for the full set of correlations and supporting details).

Yet there are notable exceptions to the rule. While the prolific cross-licensing of Intel's foundational x86 architecture—initially driven by IBM's demand for a "second source" manufacturer—helped create a powerful "innovation commons" that accelerated PC development far beyond what a single firm could have achieved, this commons was not a one-way pipeline from leader to followers. In the early 2000s, AMD introduced a pivotal improvement—extending Intel's 32-bit x86 architecture to 64-bit computing while preserving backward compatibility. The extension proved so valuable that Intel ultimately adopted the same approach, folding the follower's innovation back into the dominant platform (Shankland 2004). This "licensee-to-licensor" reversal demonstrates that cross-licensing can create a shared platform that multiple firms advance over time, with the leader retaining architectural gravity while still importing key improvements from rivals (Menaldo 2021).

THE VERTICAL NETWORK: SEMICONDUCTORS TO DEVICE MANUFACTURERS

The second network dataset maps the pre-2007 electronic digital device supply chain—a "vertical" network observing connections between upstream semiconductor firms and downstream companies including hardware manufacturers (producing PCs and mobile devices like the Palm Pilot), wireless telecommunications companies, and software publishers.[83] This network explicitly excludes horizontal ties between rivals to focus exclusively on technology flow up and down the supply chain. This matters for our diffusion story because it lets us observe not only who invents foundational components, but also who integrates them into usable systems—often the step that determines whether invention becomes mass adoption.

This network captures a supply chain in rapid evolution. The era begins in 1976 with clunky personal computers capable of basic word processing. It ends in 2006 with the industry oriented toward mobile, interconnected devices: laptops, mobile phones, and personal digital assistants with wireless capabilities. Improvements in integrated circuits and battery technology made smaller, more powerful devices possible, while modem integration connected laptops and phones to the nascent internet.

The vertical network connects 289 firms through 2,186 ties—only about 3 percent of possible connections. Low reciprocity (0.29) indicates knowledge flows are often one-way transfers from one layer of the supply chain to another. In diffusion terms, this is what we should

[83] I exclude Data Processing and Hosting (61 firms) and Computer Systems Services (206 firms, including IBM) from this network because these sectors were only indirectly involved in the core device supply chain during this era—primarily providing peripheral equipment or business-to-business services rather than components critical to the functionality of PCs, laptops, and proto-smartphones.

expect in a layered supply chain: ideas often move across layers (chips → devices → software/services), rather than being repeatedly traded back-and-forth among direct rivals within the same layer. Despite this sparseness, the network is highly efficient, with an average path length of just 2.14 and a diameter of only 5. Practically, that means an upstream breakthrough can reach a wide range of downstream applications through only a few "handoffs"—a structural precondition for rapid commercialization rather than slow, localized diffusion.

Unlike the clear hierarchy of the semiconductor network, leadership in this vertical network is distributed across sectors. The firms with highest betweenness centrality are all hardware manufacturers: Concurrent Computer Corp., Compaq, Palm, and Motorola dominate the top rankings, while Texas Instruments ranks 49th and Intel and Qualcomm have scores near zero. This suggests that in the vertical supply chain, downstream device makers acted as critical bridges connecting upstream chip innovations to the broader market. These firms are not necessarily the originators of the most foundational components; they are the system builders who recombine upstream inventions into complete products—thereby converting technical advances into diffusion at scale.

[INSERT FIGURE 5.4 HERE]

Figure 5.4: Vertical Network and the Role of Texas Instruments

Notes: This network visualization highlights Texas Instruments' connections within the vertical supply chain network of 289 firms, showing only the ties between the upstream semiconductor sector and downstream firms (hardware manufacturers, software publishers, and wireless telecommunications companies). The figure displays the 72 downstream firms whose patents cite Texas Instruments' patents between 1976 and 2006, demonstrating its role as a foundational technology provider across the digital device ecosystem. This vertical network explicitly excludes

horizontal ties between firms within the same industry sector to focus exclusively on the flow of technology up and down the supply chain. Node size is proportional to the total number of connections each firm has within the broader vertical network.

Sources: NBER Utility Patent and Patent Citation Data File (Hall et al. 2001); S&P Global Market Intelligence COMPUSTAT database (2025).

[FIGURE 5.5 HERE]

Figure 5.5: Vertical Network and the Role of Intel

Notes: This network visualization highlights Intel's connections within the vertical supply chain network of 289 firms, showing only the ties between the upstream semiconductor sector and downstream firms (hardware manufacturers, software publishers, and wireless telecommunications companies). The figure displays the 80 downstream firms whose patents cite Intel's patents between 1976 and 2006, confirming its central role as a foundational technology provider. This represents the largest downstream citation base among the "Big Three" semiconductor leaders. This vertical network explicitly excludes horizontal ties between firms within the same industry sector to focus exclusively on the flow of technology up and down the supply chain. Node size is proportional to the total number of connections each firm has within the broader vertical network.

Sources: NBER Utility Patent and Patent Citation Data File (Hall et al. 2001); S&P Global Market Intelligence COMPUSTAT database (2025).

[FIGURE 5.6 HERE]

Figure 5.6: Vertical Network and the Role of Qualcomm

Notes: This network visualization highlights Qualcomm's connections within the vertical supply chain network of 289 firms, showing only the ties between the upstream semiconductor sector

and downstream firms (hardware manufacturers, software publishers, and wireless telecommunications companies). The figure displays the 19 downstream firms whose patents cite Qualcomm's patents between 1976 and 2006, revealing its more specialized but still critical role in the supply chain. While Qualcomm's downstream citation base is smaller and more focused than Texas Instruments' or Intel's, this select group includes the same core industry titans, demonstrating that its foundational wireless technologies were essential to all major players in the digital device ecosystem. This vertical network explicitly excludes horizontal ties between firms within the same industry sector. Node size is proportional to the total number of connections each firm has within the broader vertical network.

Sources: NBER Utility Patent and Patent Citation Data File (Hall et al. 2001); S&P Global Market Intelligence COMPUSTAT database (2025).

Figures 5.4 through 5.6 visualize the direct ties from the "Big Three" semiconductor firms to the rest of the supply chain, revealing how foundational technologies from Texas Instruments, Intel, and Qualcomm diffused downstream to implementers. As Figure 5.4 reveals, Texas Instruments served as a foundational technology source for the entire downstream ecosystem. A total of 72 downstream firms cite its patents, including virtually every major technology titan of the era: telecommunications giants like AT&T, hardware pioneers like Apple and Motorola, and dominant software firms including Microsoft, Oracle, and Adobe. Figure 5.5 shows Intel was cited by an even larger group of 80 downstream firms, with high overlap among the same industry titans. Figure 5.6 reveals Qualcomm's more specialized role—cited by only 19 downstream firms, but this select group included the same core titans, demonstrating that even with fewer connections, Qualcomm's foundational wireless technologies were essential to all major players.

Downstream Leaders: AT&T, Apple, and Motorola

The vertical network highlights a central diffusion mechanism: technological progress spread fastest when upstream component innovations were integrated into complete systems and deployed at scale. The cases below show how that integration worked in practice—and why diffusion was not purely "top-down," but also included learning flowing back from system builders to component innovators.

Three firms merit particular attention for their roles in the pre-2007 digital device ecosystem. I focus on these cases not because they are the only important downstream firms, but because together they illustrate three distinct diffusion roles. AT&T shows how a "network operator" can also be an upstream knowledge source via foundational research and infrastructure; Motorola illustrates the systems-integration function that turns component technologies into working devices; and Apple exemplifies platform assembly—building a portfolio that recombines upstream innovations into a consumer-facing architecture that ultimately culminated in the smartphone era.

AT&T emerges as the most connected firm in the vertical network, with ties to 166 companies including 60 semiconductor firms that cite its patents. Its portfolio cites all major upstream players. During 1976-2006 alone, AT&T amassed nearly 12,000 patents cited over 200,000 times. The reason is Bell Labs—the legendary research laboratory that, after AT&T's 1980s breakup, retained patents for foundational electronic inventions including the transistor. AT&T's engineers also designed the Advanced Mobile Phone System (AMPS), enveloping the United States in a grid of hexagonal "cells" that connected mobile phones to the existing telephone system. A single patent, US Patent No. 4,672,658 (Gilhousen et al. 1987), illustrates the kind of foundational technology AT&T was producing. This groundbreaking patent, which

boasts a generality score of 0.84 and was cited over 240 times, describes a novel wireless local network—a direct precursor to modern Wi-Fi.

In diffusion terms, AT&T illustrates that a “downstream” firm can also be an upstream knowledge source: infrastructure and research capabilities can generate foundational inventions that many component and device makers subsequently build upon.

Motorola, the storied communications company, spearheaded the commercialization of mobile phones. In 1984 it developed the first commercial mobile phone, the DynaTAC, which used AT&T's AMPS cellular system. This iconic "Brick" phone represented Motorola's unique position as a systems integrator, combining its own radio engineering expertise with foundational network technology from AT&T and semiconductor components from upstream suppliers. Motorola also contributed essential patents to the GSM standard that marked the 2G era, enabling interoperability across national borders and introducing SIM cards.

In diffusion terms, Motorola embodies the integrator role: it converts upstream inventions into interoperable, mass-market devices, which is precisely why device makers appear as bridges in the vertical network.

Apple played a critical role in introducing personal computing to the masses with its Apple II in 1977. Several patents associated with that machine and its successors helped disseminate cheap computing capable of spreadsheets, word processing, and graphic design. Moreover, in the run-up to the iPhone's 2007 introduction, Apple built a patent portfolio drawing on technologies from key players in the PC supply chain, including wireless modem technologies from Qualcomm, digital signal processors from Texas Instruments, low-power processor architecture licensed from ARM, and power management techniques pioneered by Intel.

In diffusion terms, Apple exemplifies platform assembly: it recombines upstream components into a coherent consumer architecture, helping transform discrete inventions into broadly adopted technological systems.

Diffusion between Chipmakers and Device makers

The data reveals deep reliance of downstream device manufacturers on upstream semiconductor innovations. Apple's patents cite 235 different Texas Instruments patents and 467 different Intel patents. One of the most frequently cited Intel patents is U.S. Patent No. 5,193,167 (Sites and Witek 1993), a landmark invention for sophisticated power management—a critical technology for portable, battery-operated computers that acted as an intelligent coordinator automatically putting laptop components into low-power "sleep" states. Apple's connection to Qualcomm, while smaller (seven patents), was highly strategic, focusing on U.S. Patent No. 5,966,098 (Qi et al. 1999), a miniaturized antenna essential for small, sleek handheld devices.

Motorola's reliance was even deeper. Its patents cite a staggering 2,495 different Texas Instruments patents, over 1,000 Intel patents, and 1,241 Qualcomm patents. The most cited Qualcomm patent is the foundational CDMA invention, U.S. Patent No. 5,103,459 (Gilhousen et al. 1992), underscoring how downstream device manufacturers depended on core innovations from upstream semiconductor partners.

Importantly, innovation was not a one-way street. Leading semiconductor firms actively integrated solutions from downstream innovators, indicating feedback loops in which system-level problems and solutions (such as security, authentication, and privacy) propagated back into component design. Texas Instruments cites Motorola's portfolio 3,137 times, with particular focus on U.S. Patent No. 4,993,068 (Piosenka and Chandos 1991), an unforgeable personal identification system—suggesting upstream firms were incorporating security technologies

pioneered by hardware specialists. Intel cited Apple's patents 924 times, with particular focus on U.S. Patent No. 5,167,024 (Smith et al. 1992), Apple's patent for conserving laptop battery life, critical for developing energy-efficient mobile processors; Intel also cited Motorola 3,906 times, again focusing on the identification system patent. Qualcomm's citation behavior reflects its specialized focus on wireless communication: it shows minimal interest in Apple's portfolio (19 citations), mostly referencing power management, but it cited Motorola's patents 2,277 times, with particular focus on U.S. Patent No. 4,785,463 (Janc and Jasper 1988), a cellular privacy device patent with a generality score of .78.

Taken together, these bidirectional citation patterns show that diffusion within the high-tech ecosystem was both top-down (components to devices) and bottom-up (device-level solutions feeding back into component design). A final question is whether technological "superstars" also accelerate diffusion across the broader economy—by importing external best practices from other industries and then distributing them to their own followers.

Network Leaders as Curators?

Did industry leaders act not just as creators but as curators? The data reveals a surprising pattern: the "curator" role is far more common in non-high-tech sectors. While the average generality of backward citations for technological followers is similar across sectors (.35 for high-tech vs. .38 for non-high-tech), the difference for leaders is stark. The average for high-tech leaders is only .39, while for non-high-tech leaders it is .56. Leaders in industries like chemical and vehicle manufacturing are significantly more open to citing patents from outside their own fields (see Menaldo 2026, Section 2 for additional examples).

This seemingly surprising finding offers an important insight into the nature of superstar economics. The relative insularity of high-tech leaders isn't a weakness; it's a defining

characteristic of their dominance during the Third Industrial Revolution. These firms weren't just participating in the economy; they were creating a new one. Rather than importing best practices from other sectors, they were the best practice—setting the technological standards and codifying processes that underpinned the transition to the digital economy. In diffusion terms, their relative insularity is the flip side of their hub status: they did not need to import external general-purpose technologies because they were generating the bedrock technologies around which the rest of the ecosystem coordinated.

CONCLUSION

The previous chapter charted a profound economic transformation, tracing the arc from postwar globalization to a U.S. economy built on intangible capital; the statistical evidence presented in this chapter demonstrates that this transformation was neither incremental nor evenly distributed. High-tech sectors exhibited a Pareto-like "winner-take-most" pattern: the top firms accounted for the overwhelming share of forward patent citations, and firms at the technological frontier displayed materially higher productivity than their peers. Patent distributions followed a power-law dynamic distinctive to high-tech industries, reflecting a sharply concentrated innovation frontier.

Network analysis then identifies the mechanism through which this concentrated frontier nonetheless generated broad diffusion. In the horizontal semiconductor network, Texas Instruments, Intel, and Qualcomm emerge as central hubs through which general-purpose technologies propagated to a wide periphery of adopters. Yet this asymmetry does not imply technological stagnation or monopolization: the networks reveal a dense but open innovation commons in which foundational platform technologies spread quickly through short paths and are subsequently refined, implemented, and extended by followers—avoiding the "dual

economy" pattern observed in Japan, where frontier innovations remained more internally contained.

The vertical network clarifies a complementary diffusion channel: commercialization depends on system builders that integrate upstream component breakthroughs into complete devices and deploy them at scale. Downstream firms appear as key bridges precisely because integration converts invention into adoption. The patent-level evidence makes these channels tangible. Downstream firms such as Apple and Motorola cite upstream inventions central to portable computing and mobile communications—ranging from sophisticated power management to foundational wireless technologies such as CDMA and miniaturized antennas. Importantly, diffusion is not one-way: upstream leaders also cite downstream innovators (e.g., Texas Instruments' heavy citation of Motorola's authentication and security inventions), indicating feedback loops in which device-level problems and solutions propagate back into component design.

Finally, the chapter's "curator" evidence highlights an important boundary condition. High-tech leaders were extraordinarily effective at diffusing their own general-purpose technologies within the ecosystem, but they were relatively insular compared to leaders in non-high-tech sectors in adopting ideas from outside their fields. This insularity was not a weakness but a defining feature of their dominance: rather than importing best practices, they became the best practice—generating the bedrock technologies around which the rest of the high-tech ecosystem coordinated.

Yet as powerful as this decentralized diffusion mechanism proved, it had limits. While it effectively propagated foundational technologies within and across layers of high-tech supply chains, it could not by itself guarantee the explicit global interoperability required for the next

leap: the truly mobile internet. That challenge demanded a different model of collaboration—formal coordination through Standards-Setting Organizations and licensing of essential patents on FRAND terms. The next chapter turns to that transition through the case of Qualcomm and the development of 4G mobile technology, where institutionalized standard setting became indispensable for building the globally interconnected world that would give rise to the smartphone—and, ultimately, the data-rich foundation of the Fourth Industrial Revolution.

CHAPTER 6: SMARTPHONE SUPPLY CHAIN & INTEROPERABILITY

During the late 1990s, Netscape's free distribution of its Navigator browser—paired with enterprise licensing—popularized, at browser scale, a freemium model that lowered adoption costs and drove demand for internet bandwidth. The Web's open, standards-based architecture—rooted in protocols and institutions such as HTTP/HTML, TCP/IP, DNS, the W3C, and the IETF—stood in contrast to closed, curated services like AOL's walled garden, accelerating adoption and catalyzing a wave of commercial applications.[84] In short order, nominally free-to-user services—search, e-commerce, online classifieds, and social media—reorganized markets and undermined many incumbent media, retail, and telecommunications business models. The emerging ecosystem soon eclipsed the browser that helped launch it: in 1998–1999 AOL

[84] The early internet was built on a series of foundational standards and protocols to ensure universal interoperability. At its core is TCP/IP (Transmission Control Protocol/Internet Protocol), the basic language that dictates how data is broken into packets, sent to the correct destination, and reassembled in the right order. This system is complemented by HTTP (Hypertext Transfer Protocol), the protocol for sending and receiving web pages, and HTML (Hypertext Markup Language), the language used to create those pages, which together make the World Wide Web work. To make navigation user-friendly, the DNS (Domain Name System) acts like the internet's phonebook, translating human-readable addresses (like apple.com) into the numerical addresses computers need. At the time of this writing, these open standards and protocols were still developed and maintained by key organizations such as the W3C (World Wide Web Consortium) and the IETF (Internet Engineering Task Force), ensuring the internet remained an open and collaborative platform. On all of this see Kurose and Ross 2021.

acquired Netscape for approximately $4.2 billion, marking a swift reversal from Netscape's mid-1990s prominence, epitomized when the stock more than doubled on the day of its initial public offering.

The explosion of the open web created an insatiable demand for access—which led computer makers to scramble to reduce the cost of computing hardware, including storage. In hindsight, the "openness" of the web (protocols) logically entailed the "openness" of the hardware supply chain (specialization) to scale. Smartphone production networks crystallized this logic, with standardized interfaces and modular components enabling globally distributed manufacturing and rapid iteration. But, of course, the demand for innovations of this sort does not automatically translate into its supply.

To confidently engage in constant R&D and focus exclusively on design while outsourcing the tangible portions of the chip supply chain, upstream firms needed strong safeguards to their IP. Consider that a firm that purchases or rents an idea cannot fully assess its value until its details are revealed; but, once it's revealed, that information is effectively transferred, making it difficult for the seller to prevent unauthorized use or appropriation (Arrow 1962). As Chapter 2 made clear, strong IPRs provide necessary excludability, turning non-rivalrous information into excludable assets, thus incentivizing disclosure and commercialization in a specialized, multi-firm ecosystem (see Barnett 2020; Teece 1982). They also reduce the transaction costs that affect inventors' search for partners who can help them commercialize their innovations, including input providers and retailers; IP underpins the web of contracts that brings new supply chains into being (see Spulber 2015; Haber 2016). This is especially the case for semiconductors (Galetovic 2021).

And as I outlined in Chapter 3, US IP policy underwent significant changes during the CDP era precisely to coax novel innovation ecosystems into being. This included endowing patent enforcement greater coherence through the creation of a specialized IP court, the CAFC; explicitly recognizing software patents; and introducing a framework for commercializing federally funded research undertaken in US universities through the Bayh-Dole Act. These reforms, alongside the globalization of US-style IP protections via TRIPS and similar efforts attached to multilateral trade agreements, fundamentally enabled upstream firms to confidently concentrate on intangible asset creation, trusting that their innovations would be protected across borders and through complex contractual relationships.

By the early 2010s, Apple leveraged this framework to master the orchestration of a globally distributed smartphone production network that was later emulated by its rivals. Final assembly was handled by contract manufacturers such as Foxconn and Pegatron, while a silicon stack combined in-house design with specialist suppliers hired by contract.[85] Apple designed its

[85] Like in a computer, a mobile phone chip functions through transistors. To review and simplify what Chapter 2 explained: these tiny pieces of hardware are essentially glorified switches. When a transistor is on, it represents a one, and when it's off, it represents a zero. Ultimately, all computer programming is translated into a series of ones and zeros that the processor understands, so processors can be thought of as engines that continuously switch between ones and zeros based on what program they're fed (see Petzold 2022). The complexity of the underlying chip changes based on the complexity of the processes it needs to handle. Each processor itself contains multiple components. These include the processing unit, which consists of the transistors—arrayed and connected within integrated circuits on wafers of silicon; the

own application processors (A-series system-on-chip designs) and sourced cellular baseband modems featured on those System on Chips (SoCs) from external vendors such as Qualcomm.[86] Other functions—Wi-Fi/Bluetooth, sensors, memory, and power management—came from dedicated component suppliers.[87]

---

control unit, which functions like a traffic signal—telling the program or incoming data where it must go and coordinating what's happening on the processor; and memory, where pieces of information are stored before being fed to the processor to perform its operations (Patterson and Hennessy 2021).

[86] As I will elucidate at greater length further below, SoC is a type of integrated circuit that combines nearly all the essential components of a computer—such as the central processor, graphics processor, memory, and wireless modem—onto a single chip. This contrasts with traditional designs that used separate chips for each function. The main advantage of an SoC is its ability to be more compact, faster, and far more energy-efficient, which is critical for devices with limited space and battery life like smartphones.

[87] While DRAM and NAND flash memory are critical semiconductor components in every handset, they sit in a somewhat distinct supply chain: they are typically designed and fabricated by specialized memory producers such as Samsung, SK Hynix, and Micron, follow their own scaling trajectories, and are often purchased as separate components that are integrated alongside the SoC at the device-assembly stage. Therefore, although I will not explore these further, I note that memory chips are highly capital-intensive and cyclical, often approaching commodity-like pricing.

Importantly, the supply chain for the chips themselves was extremely vertically disintegrated, sophisticated, and global. While I will briefly outline a smartphone's full bill of materials to properly scaffold the argument, this chapter will focus on the ARM-centric SoC pipeline—IP licensing, logic design, foundry processes, EUV lithography, and advanced packaging to overcome scaling limits. Upstream, fabless chip firms—including Apple, Qualcomm, Broadcom, and (for some Android devices) Nvidia—relied on electronic design-automation (EDA) tools from Synopsys, Cadence, and Mentor (now Siemens EDA) to simulate, verify, and lay out chips with billions of transistors. The chips themselves were manufactured in Asia by foundries like TSMC and Samsung; they were the only ones with the sci-fi-like facilities needed to make semiconductor wafers at large scales. Yet, to fabricate chips at resolutions 10,000 times finer than human hair, these semiconductor fabs imported specialized machines from a Dutch firm, ASML, alongside critical etching and deposition tools from US giants like Applied Materials and Lam Research and niche firms in Japan.

Players operating across this supply chain mastered business models that reflected the logic of specialization along the lines of comparative advantage (Barnett 2020). Consider that while initially Qualcomm manufactured its own handsets in the US, this was a strategic move solely to demonstrate the viability of its Code Division Multiple Access (CDMA) technology, an essential ingredient to modern smartphones that I will discuss in greater detail further below. Once CDMA's value was proven, Qualcomm swiftly divested from phone and equipment manufacturing, choosing instead to concentrate its efforts on core technology development and chip design (Svensson 2024). Similarly, InterDigital never ventured into phone manufacturing either. Much like Qualcomm, it had amassed one of the largest portfolios of wireless SEPs by 2025, and generated revenue exclusively through licensing (see Galetovic et al. 2017). And a

multitude of smaller firms that also specialized in developing critical, invisible technologies inside smartphones licensed their innovations through patent pools, allowing them to circumvent the significant overhead of maintaining IP divisions (see Via Licensing Corporation 2018).

To throw this into sharp relief, consider that before the rise of fabless designers and disaggregated global supply chains for smartphones and other mobile devices, the semiconductor industry was dominated by vertically integrated giants that managed every stage of production under one roof because the technology was still relatively simple and the industry's economics favored this approach (Mazurek 1999) During the 1980s, leading Japanese firms like NEC, Toshiba, and Hitachi not only designed and developed chip architectures, but also fabricated wafers in their own fabs, performed assembly and testing operations, and handled global sales and marketing. By 1990, Japanese companies commanded roughly 50 percent of the world's semiconductor market: they supplied high-reliability memory modules for mainframes with long product lifecycles backed by 25-year warranties.

Similarly, American-based Intel perfected a vertically integrated model centered on the burgeoning PC market. The company combined in-house microprocessor design (see Chapter 2) and software tool development with fully owned wafer fabs, packaging, test facilities, and a global sales force.[88] Intel's comprehensive control over the chipmaking process allowed it to seamlessly integrate manufacturing improvements—from shrinking features and adopting new

[88] Streaming SIMD Extensions (SSE) allow CPUs to efficiently handle several numbers simultaneously. For example, a processor can add four pairs of 32-bit floating-point values—numbers with decimal points that a computer stores using 32 binary digits (bits), allowing for a wide range of very precise or very large/small numbers (Patterson and Hennessy 2021).

materials to refining transistor wiring—with architecture optimizations that did not require changing how the semiconductors were physically made. Consequently, high levels of R&D paid off: every time Intel crammed more transistors into its chips, this directly translated into tangible performance gains for PC users who in turn responded by boosting its sales (see Galetovic 2021).

However, in the face of accelerated product cycles driven by relentless technological progress around semiconductors, cellular networks, smartphones, the app economy, and AI, pursuing an industrial organization strategy based on vertical integration within a single nation became a fool's errand (see Kessler 2022; Kessler 2025). Consider that Intel's vertically integrated model proved fundamentally maladapted to the smartphone age, let alone the subsequent AI revolution. While the company spent decades pushing ever-faster chips with power consumption as a secondary concern, it failed to develop a competitive mobile processor in the 2000s, losing a market worth hundreds of billions to energy-efficient, ARM-based processors designed for the low-energy demands critical for battery-powered mobile devices (see Gilbert and Rosenthal 2024).

To boot, this rigidity caught up to Intel not only in mobile, but also in PCs, as illustrated when Apple transitioned its Macintosh line to its own ARM-based M-series processors, offering orders of magnitude improvements in both processing power and battery life. Indeed, by 2025, Microsoft had also embraced ARM-based Snapdragon CPUs designed by Qualcomm for its Surface tablets. And Intel's woes even extended to data centers: as energy consumption became a significant expense during the early 2020s, ARM designs gained a steady following.

By the mid-2020s, even Intel implicitly conceded the limits of the traditional vertically integrated model. Under its 'IDM 2.0' strategy, the company began opening its factory doors to manufacture chips for outside clients—effectively becoming a foundry itself—and launched

massive, government-backed fab projects in Ohio and Germany to compete directly with TSMC in the contract manufacturing game (see Goswami and Leswing 2024).[89]

Attempts to re-integrate production on domestic soil have fared poorly. Purism's "Liberty Phone"—assembled in California with a motherboard built in-house—sold fewer than 100,000 units in 2024 compared to Apple's 225 million, at higher prices and inferior performance (Raab 2025).[90] The lesson: sub-scale domestic operations cannot replicate the deep pools of talent and capital distributed across global supply chains, and some analysts project a shortage of roughly 67,000 technicians and engineers in the U.S. semiconductor sector by 2030 (Wiseman et al. 2025).

---

[89] Because no good deed goes unpunished, after Intel laid off thousands, cut its dividend, and attempted to spin off its foundry services to fulfill contract orders for fabless clients, it was beset by production delays and abandoned acquisitions; plus, a surge in AI workloads favored GPU chips like those designed by Nvidia and fabricated by TSMC, not CPUs (see AInvest 2025).

[90] The NXP i.MX 8M Quad is built on a mature 28 nm process (planar) node, whereas leading mobile application processors use 5 nm or even 3 nm technology that employs a 'fin-style' transistor layout—essentially wrapping the channel around a vertical fin or gate—that lets chipmakers cram far more switches into the same area and cut power use in half. This 3D arrangement is what enables today's mobile chips to deliver huge performance gains and battery life improvements over a planar 28 nm design (see Techlevated 2023). This means that compared with modern chips fabricated on denser, more advanced nodes, the Liberty's chip has fewer larger transistors that deliver lower peak performance and higher power consumption.

Finally, beyond vertical disintegration as such, as I outlined in Chapter 3, and will take up again in this chapter, SSOs were also critical for coordinating high levels of technological specialization split between dozens and dozens of firms.[91] As readers may recall from Chapter 3, these organizations convene representatives from various companies and industries to develop standards aimed at achieving specific goals or solving problems across consumer goods, manufacturing, and commercial activities. These objectives can range from ensuring product quality—for example, a standard like International Organization for Standardization 9001, which specifies requirements for a quality management system relevant to organizations of all types and sizes—or establishing technical specifications for materials and products, such as ASTM

[91] As I will touch on in Chapter 8, vertical disintegration is also embodied in the software that runs mobile devices. In the case of Android smartphones, manufacturers rely on Google's Android operating system, which they may customize but did not develop from scratch. Instead of a single, closed hardware-software system like Apple's, Google offered its core "Android Open Source Project" for free to any hardware manufacturer. This commoditized the base operating system, sparking a horizontally competitive hardware market where dozens of companies, from Samsung to Xiaomi, competed on features and price. However, to access the crucial, proprietary layer—Google Mobile Services (GMS), which includes the essential Play Store, Maps, and Search—manufacturers had to agree to Google's licensing terms. This strategy allowed Google to maintain vertical control over the software and services stack that fuels its data flywheel while outsourcing the immense cost and risk of hardware manufacturing to a vast ecosystem of partners (see Campbell-Kelly et al. 2015).

International D638, which provides a standard test method for evaluating the tensile properties of plastics used in manufacturing (Ramkumar 2025).

In the case of smartphones, SSOs evolved into an essential "governance layer" during the latter part of the Third Industrial Revolution, simulating some of the functions of the Chamberlian corporate hierarchy that characterized the 20th century firm. Whereas vertically integrated giants like AT&T or Xerox once used middle managers and research labs to coordinate compatibility between their internal divisions and stimulate innovation, the disintegrated smartphone supply chain came to rely on SSOs to coordinate technical standards across hundreds of independent firms. In this new industrial model, consensus-based protocols, rather than executive fiat, catalyzed innovation within a heterogenous global production network. This chapter will illustrate how SSOs defined connectivity protocols for systems and processes to ensure digital devices could seamlessly connect and efficiently operate with other devices, regardless of their manufacturer.[92] It will also explore their unique industrial organization. Ahead I will argue that the SSO story is somewhat complicated, however. They are necessary but not sufficient for undergirding complex and value enhancing fractionalized supply chains: despite declaring their patents essential to technical standards, upstream firms with massive sunk R&D costs may still face some appropriation by downstream implementers who can leverage

[92] Interoperability refers to a common language or infrastructure between products, devices and networks that allows them to work with each other. This principle is readily observed in everyday standards, such as the electrical outlets and compatible plugs in our homes, the standardized lightbulb sockets designed to work with all bulb types, and the Universal Serial Bus (USB) ports on computers that accommodate a wide array of peripheral devices.

litigation, regulatory pressure, and the ambiguity of FRAND commitments to renegotiate terms ex post.

This chapter continues as follows. I first trace how the smartphone became the emblematic product of a vertically disintegrated, globally distributed innovation system—and why that system required not only modular engineering, but also a governance layer that could sustain massive sunk R&D. I map the modern smartphone supply chain and then zoom in on the ARM-centric SoC pipeline, moving from IP licensing and chip design to EDA software, foundry fabrication, lithography equipment, and advanced packaging—each segment characterized by distinct cost structures, scale economies, and strategic bottlenecks. Next, I show how this fragmented production network depended on interoperability standards, focusing on the 4G LTE regime that turned the smartphone into an always-connected computer and unlocked the app economy. I then explore the institutional machinery that made such coordination possible: standards-setting organizations, standard-essential patents, and FRAND commitments, which jointly aim to balance upstream appropriation incentives against downstream implementation. Finally, I use Qualcomm as a central case to illustrate how leadership in SSOs and SEP portfolios can enable innovation while sharpening bargaining conflict—clarifying why, in practice, standards are necessary but not sufficient to prevent opportunism in complex supply chains.

THE MOBILE PHONE SUPPLY CHAIN

The evolution of mobile phones exemplifies the culmination of vertical disintegration in supply chains.[93] The smartphone stack is characterized by distinct layers of specialization: chip designers, chipmakers, handset manufacturers, operating system providers, and a vast ecosystem of third-party application developers. During the CDP era, the assembly of these devices, including Apple's iPhones and other mobile digital devices, such as Motorola and Sony phones running the Android platform, became largely separated from the manufacturing of semiconductors and other key components like Gorilla Glass.

Indeed, before the iPhone became the mobile phone standard, Motorola proactively outsourced much of its manufacturing to Flextronics and other contract manufacturers in the 2000s. This strategic move allowed Motorola to shed the burden of massive capital expenditures (CAPEX) for manufacturing facilities, instead focusing its resources on product design, brand development, and marketing. It represented a precocious recognition that specializing in core strengths and offloading capital-intensive production could lead to greater agility and efficiency in a rapidly evolving market (Serant 2000).

Beginning in the late aughts, Apple had taken the outsourcing of mobile phone component production to the next level. The company contracted with Foxconn and Pegatron to assemble these devices. Corning supplied Gorilla Glass; Sony manufactured camera sensors; Qualcomm designed processors and modems; and TSMC fabricated custom-designed chips (Pineda 2021; Elmore 2024).

[93] To be more exact, this refers to the post-2010 era, as Motorola was originally known for its own proprietary OS before pivoting to Android.

The underlying logic was that by decoupling high-margin design and software innovation from capital-intensive manufacturing, the firm could achieve infinite scalability and financial efficiency, leveraging the brutal competition among suppliers to drive down costs while reserving its own capital for the high-leverage activities that differentiated its products (Dedrick et al. 2010).[94] While designing a mobile digital device carried prohibitive up-front R&D outlays, key components, such as its modem chips, became highly interoperable and modular. Meanwhile, smartphone fabrication and assembly entailed massive, sunk capital investments in large facilities that promised only razor-thin per-unit returns. Specialization promised handsome returns in the face of these changing economics, especially to the idea-intensive firms operating upstream, as it allowed them to preserve their balance sheets for intangible, design-driven investments that allow them to truly differentiate their products. It also allowed them to have their cake and eat it too: they could flexibly contract with manufacturers that gave them instant access to world-class, state-of-the-art capacity and could surge production as needed to bring new releases to market (Barnett 2020).

[94] In the early 2020s, gross profit margins for mass-market smartphone vendors were often in the low double digits or even single digits. For example, Xiaomi's gross margin on smartphones in early 2024 was just 13.5%—an average gross profit of about $21 per device on an average revenue of $156 per device. For low- and mid-range devices, profit margins were even slimmer, sometimes below 10% or close to breakeven. While high-end models such as the iPhone Pro Max could achieve gross profit margins above 50%, these were exceptions that proved the rule: volume often mattered more than per-unit profitability (see Bjorhovde 2024; Galetovic 2021; Isaac 2025).

Indeed, a brief, but significant, departure from the vertical disintegration model illustrates its powerful attraction. Google itself experimented with vertical integration by acquiring handset maker Motorola Mobility in 2012, a move driven by a desire to acquire Motorola's vast patent portfolio and to create its own "stock" Android devices. However, the experiment was short-lived; by 2014, Google had sold the handset business to Lenovo, choosing to revert to its disintegrated model of providing software to a wide range of hardware partners while still retaining the valuable IP (see Page 2014).[95]

Returning to Apple, the highly disintegrated digital device model it pioneered allowed it to concentrate on design, software development, and marketing. In turn, it scaled rapidly and maintained its focus on R&D and the user experience, rather than tying up capital in manufacturing plants (Yan et al. 2022). Indeed, Apple even outsourced a lot of more technical oriented innovations under this paradigm. Incredibly, its R&D intensity as a share of revenue was generally lower than that of upstream technology licensors; for example, in 2024, while Apple spent about 8% of revenue on R&D, Qualcomm, whose sole focus is on core

[95] While Google stepped back from manufacturing mass-market handsets for a time, it later re-entered the vertical integration game with its Pixel line of smartphones. Unlike the Motorola acquisition, which was largely about acquiring patents, the Pixel strategy is centered on tight integration between hardware and software to optimize AI performance—most notably through Google's proprietary Tensor SoC designed specifically to accelerate machine learning tasks on the device (see Chapter 9).

technological innovation, spending 23% of its revenues on R&D.[96] This difference underscores how contract-manufactured assembly was only the first wave of vertical disintegration in mobile phones, the next one occurred within the chips themselves.

SMARTPHONES' SEMICONDUCTOR SUPPLY CHAIN

In the chapter's introduction, I briefly discussed how, during the late 20th and early 21st centuries, a vertically disintegrated supply chain for advanced chips displaced integrated semiconductor giants (Barnett 2020; Galetovic et al. 2017; Galetovic et al. 2024; Macher and Mowery 2004). Fabless chip design, EDA software creation, the manufacturing of integrated circuit etching equipment, chip fabrication, semiconductor testing, and chip packaging became distinct processes. Each was eventually handled by highly specialized firms operating in different countries according to their competitive advantage.[97] Companies like Qualcomm, NVIDIA, and AMD honed chip design expertise concentrated in the US. Actual chip production overwhelmingly shifted overseas to specialized foundries in East Asia, notably TSMC in Taiwan

---

[96] That said, in 2025 Apple committed to significantly expanding its R&D investment across AI, custom chips, and software innovations (Apple Inc. 2025). I calculated these numbers using Alsop (2025) and Laricchia (2024).

[97] Companies such as Samsung operate dedicated factories that produce specific types of memory chips: DRAM (Dynamic Random-Access Memory). These provide a computer's fast, temporary working memory, and Not And (NAND) flash memory, which is used for long-term storage (Patterson and Hennessy 2021). These specialized chips also follow Moore's Law, as they consistently shrunk and became more powerful over several decades (see Figure 2.1 in Chapter 2).

for logic chips and specialized semiconductors, and South Korea for both memory and advanced logic (Samsung), and Japan for specialized components and memory. Meanwhile, the firms that create highly sophisticated EDA software used by chip designers, including Synopsys and Cadence Design Systems, were headquartered in the US. ASML in the Netherlands mastered the process of fabricating the massive, yet extremely delicate, lithographic machines that etch integrated circuits onto chips.

Revisiting Moore's Law

Readers may recall that the exponential improvements associated with chips since the early 1970s is not merely a story about increasing transistor counts; rather, it was the byproduct of semiconductor firms' relentless, creative R&D efforts to shrink chip sizes, bolster efficiency, and reduce costs (Galetovic 2021). Over the course of five decades, feature sizes plummeted from the several-micron scale of the 1970s to sub-2 nm nodes—a roughly 5,000-fold miniaturization.

As chips became incredibly tiny—below 500 nm in the 1990s and then under 100 nm in the 2000s—chipmakers faced escalating problems, including resistance, leakage, and heat dissipation. Beyond power, manufacturing also faced a critical hurdle: yields—the percentage of defect-free dies per wafer—initially suffered at each new node, forcing chipmakers to climb steep learning curves to recover throughput.[98]

[98] Historically, successive "process nodes" in semiconductor manufacturing were associated with the minimum feature size (on the order of nanometers) that can be patterned on a chip (see Sze and Ng 2006). Today, while still denoting generation-over-generation advances in transistor density and performance, the term "node" often serves as a commercial standard rather than an exact physical measurement of the chip's smallest features (Miller 2022).

To address these constraints, the industry introduced copper interconnects, high-κ/metal-gate dielectrics, and multi-gate transistor architectures such as FinFETs—advances made possible by deep-ultraviolet (DUV) lithography.[99] As the Third Industrial Revolution gave way to the Fourth, leading-edge chipmaking pushed past the 10 nm barrier using extreme ultraviolet (EUV) lithography and next-generation gate-all-around nanosheet transistors, extending density gains while controlling leakage and power.

Progress also increasingly came from "beyond scaling" techniques. Backside power delivery (BPD) moves power lines to the underside of the chip to reduce energy loss and free frontside routing capacity (Lin et al. 2022).[100] And 3D stacking and advanced packaging shorten

---

[99] In a transistor, the control gate is the crucial electrode that regulates the flow of electrical current through the transistor's main conducting channel (the path between the "source" where current enters and the "drain" where it exits). By applying a voltage to the gate, an electric field is created which acts as a switch, either allowing current to flow (turning the transistor "on") or blocking it (turning it "off"). In FinFETs, the gate's ability to wrap around the fin-shaped channel provides significantly enhanced electrostatic control over this current flow compared to older, flatter designs. This improved control is vital for minimizing power leakage and maximizing current flow, especially as transistors shrink to atomic scales (see Zhang 2022).

[100] To understand why this matters, think of a traditional chip as a high-tech skyscraper where the power lines, data cables, and plumbing are all crammed into the same vertical shafts. This creates "congestion," limiting the space available for the data cables and causing signal interference. BPD is much like building a dedicated power utility conduit on the outside of the skyscraper. By moving the power delivery network to the back of the chip, it frees up all the internal space for

the distance between logic and memory by integrating components vertically, boosting speed and energy efficiency (Zhu et al. 2013).

Having sketched the technological and economic logic of this trajectory, I now turn to the specialized segments of the smartphone semiconductor supply chain—starting with fabless chip designers and the IP they license, and then moving through EDA, foundries, lithography, and advanced packaging.

Fabless Chip Designers

Firms that outsourced chip manufacturing called on both constant innovation and marketing prowess to secure foundry slots in a market where capacity was rationed and scheduling was fierce. While the fabless chip design segment of the smartphone semiconductor supply chain evolved in the crucible of generous amounts of intangible R&D and learning by doing across architectures, IP libraries, and design software, venture funding injected capital into promising designs because winning a single tape-out at a leading node could yield exceptionally high profit margins (Miller 2022). In turn, this depended on sometimes creative IP licensing regimes—which themselves depended on the type of stronger IP enforcement outlined in Chapter 3.

Among fabless chip designers, ARM stands out. For the sake of both reducing their R&D costs and interoperability, most chip designers licensed processor instruction sets (LPIS) from U.K.-based ARM.[101] In essence, the LPIS is a standardized "language" that allows a wide range

---

more and faster signal pathways, dramatically improving performance and power efficiency (see Skibba et al 2015 for a similar analogy).

[101] More technically, a LPIS is the core set of commands that a computer chip is designed to understand and execute.

of different companies to build customized chips while still participating in a shared digital ecosystem. For example, a software developer can create an application for Google's Android operating system that can also run seamlessly on a Samsung phone with a Qualcomm Snapdragon chip and an entirely different phone with a MediaTek chip—both processors are built on the same ARM instruction set. Similarly, though Apple designs its own unique chips and software stack (iOS), it also builds its processors on the ARM instruction set; this shared foundation is what enabled Apple to unify its iPhone, iPad, and Mac platforms under a common architecture, allowing developers to write apps that run across all Apple devices with minimal modification (see Gilbert and Rosenthal 2024).

By 2025, its processor cores powered 95 percent or more of all smartphones as chip designers licensed the British company's architectures or pre-designed cores as components in their own designs (Arm Holdings plc 2023).[102] The company, originating in 1983 as Acorn's RISC (Reduced Instruction Set Computing) project, built its core philosophy around creating simpler, more energy-efficient CPUs compared to complex CISC (Complex Instruction Set Computing) architectures (Hennessy and Patterson 2019). While this focus on energy efficiency proved critical for battery-powered mobile devices, allowing for significantly extended battery life, ARM processors rapidly evolved to deliver impressive performance, which was critical for modern smartphones' multitasking and computational demands (ibid).

[102] By 2025, ARM achieved near-universal adoption: over 250 billion ARM-based chips had been shipped, with ARM's architecture present in over 99% of smartphones globally (Karlsson 2023).

A key to ARM's pervasive influence during the smartphone era was its licensing model, which allowed a vast ecosystem of companies—from giants like Apple, Samsung, and Qualcomm to smaller manufacturers—to design their own highly customized processors based on ARM's foundational architecture (Gilbert and Rosenthal 2024). This customization and flexibility spurred a diverse range of optimized chips, fostering intense innovation and integration of multiple functionalities into powerful SoCs (ibid). This, in turn, underpinned a robust ecosystem of software, tools, and developer support for Android and iOS, accelerating app development and consumer adoption. Consider Apple's A-series processors, which helped make mobile computing devices faster, more compact, and capable of fanless designs due to their lower heat generation (Apple Inc. 2023b). Or take Qualcomm's Snapdragon SoCs, which were widely adopted in Android smartphones because they combined the CPU (based on ARM architecture), GPU, modem, and other essential components to optimize performance and reduce power consumption (Qualcomm Inc. 2023).

In addition to innovative SoC designs, Qualcomm established itself as a leader in modem technology and wireless connectivity during the smartphone era for both Android and iOS devices. As I will explore in detail further below, the company provided modem chips that supported various cellular standards, with Fourth Generation Long-term Evolution (4G LTE) as the most important.[103]

---

[103] Beyond core wireless technologies around wireless connectivity such as 4G (cellular), Wi-Fi, and Bluetooth, open standards and protocols underpin nearly every one of the iphone's functions (Apple Inc. 2024b), as well as that of other smartphones. They enable universal interoperability with networks, applications, and devices These include foundational internet standards such as

True to its upstream location in the smartphone chip supply chain, Qualcomm accumulated an extensive portfolio of patents related to wireless communication technologies over several decades. By 2022, Qualcomm's worldwide patent portfolio included over 140,000 patents and patent applications, nearly half of which—some 65,000—were disclosed as cellular SEPs (Qualcomm Inc. 2022a). The company developed and owns CDMA, the foundational invention by which multiple data streams can use the same set of radio frequencies. Moreover, circa 2025 Qualcomm's patents were primarily implemented within its modem chip, rather than the broader features of a handset, with many of its essential patents being specific to communication-related components (Qualcomm Inc. 2019a).

During the early 2020s, the company's healthy financials reflected its distinctive business model.[104] Qualcomm generated $38.96 billion in total revenue in fiscal 2024, of which $33.2 billion (85 percent) came from Qualcomm CDMA Technologies, its chip sales for handsets, cars, and IoT. Qualcomm Technology Licensing—its patent-royalty arm—brought in $5.57 billion (15 percent of sales), and, thanks to pretax margins above 70 percent, contributed the lion's share of the company's profits. To achieve this enviable market position, Qualcomm put a premium on

---

TCP/IP and HTTP/HTTPS for all online communication and web browsing. Contactless payment features are made possible by the NFC protocol, while location services draw on the broader GNSS (Global Navigation Satellite System). The iphone also leverages specific standards for sending and receiving emails (IMAP/POP3), displaying images and playing media files (JPEG, PNG, MP4), and fast charging (USB Power Delivery).

[104] All the facts and figures in this paragraph are from Qualcomm Inc. 2025.

innovation: in both 2024 and 2025 it invested almost $10 billion annually in R&D to enable handsets to be "smart"—to connect at ever-increasing speeds and reliability to cellular networks.

EDA Software

During the mobile phone revolution, fabless chip designer success stories like Qualcomm's were made possible by ever-more-powerful EDA tools (Miller 2022). EDA refers to a category of specialized software tools used by chip designers. Because modern chips contain billions of transistors, designing them by hand is impossible. EDA tools automate and streamline this process, allowing engineers to create logical representations of circuits and physically arrange components on silicon chips; run virtual tests to predict circuit behavior under various conditions; and verify that designs adhere to the thousands of manufacturing rules set by foundries (see Galetovic 2021). Without EDA software, the creation of modern, high-performance processors for smartphones would simply not be possible.

By 2025, this segment was defined by exceptionally high R&D costs; leading firms continuously reinvested over 30 percent of their revenues into R&D to create sophisticated algorithms capable of simulating, verifying, and optimizing chip designs with ever-increasing complexity (Smith 2024). Circa 2025, the market for these indispensable tools exhibited relatively high levels of concentration, with the top four firms controlling 90 percent of market share globally due to immense technical barriers to entry and high switching costs (ibid). This strong market dominance allowed EDA software providers to exercise significant pricing leverage over chip designers (Mordor Intelligence 2024).

Chip Fabs

By the late 2020s, TSMC served a broad range of customers in the smartphone industry, including major smartphone chip designers such as Apple, Qualcomm, and MediaTek, enabling

it to produce cutting-edge chips that power a wide variety of devices across different price points and market segments (Miller 2022). It and other chip foundries faced the relentless pursuit of miniaturization demanded by the desire to extend Moore's Law combined with the technical exigencies of new technologies like AI and 5G.[105]

TSMC and its ilk therefore faced quite brutal microeconomics. Chip foundries were faced with the need to make constant investments in new advanced materials and techniques (Caines et al. 2025). These, however, carried incredibly high fixed and sunk costs associated with building and maintaining state-of-the-art manufacturing facilities and, simultaneously, rapidly depreciating hardware due to the high velocity of technological progress.[106] For instance, producing a leading-edge 3nm chip requires up to 110 mask layers, which is nearly triple the amount needed for older nodes and 60 percent more materials (Wiseman et al. 2025). Meanwhile, the R&D investment required to develop the manufacturing process that fabricated this chip reached approximately $5 billion (Lin 2023).

[105] Although by 2025 marketing labels no longer mapped perfectly to a single physical dimension, smaller-numbered nodes generally correspond to higher transistor density and more advanced process technologies, which in turn enable designers to pack more transistors into a given area and improve performance and energy efficiency at the chip and system level (see Miller 2022).

[106] The R&D cost to design the cutting-edge 2nm chip exceeded $700 million (Caines et al. 2025).

To illustrate this phenomenon dynamically, consider Figure 6.1, which depicts nearly exponential growth in CAPEX and R&D for TSMC over the early 21st century.[107] By 2022, TSMC's CAPEX exceeded $33 billion—up from under $5 billion in 2010—while its R&D spending surpassed $4 billion. This blowout spending allowed the company to remain at the forefront of semiconductor manufacturing technology across several generations of process nodes, including 7nm, 5nm, and 3nm—with each subsequent advance representing higher transistor density that translates to more compact chip designs, improved performance, and lower power consumption (TSMC 2023a). In turn, newer chip vintages translated into appreciable improvements in smartphone design, culminating in devices that could run more demanding applications while limiting heat and preserving battery life (TSMC 2023b).

[FIGURE 6.1 HERE]

Figure 6.1: The Costs of Running a Chip Foundry—TSMC's Escalating Expenses.

Notes: R&D Expenses include employee salaries and benefits, materials and supplies, equipment and facilities depreciation, external services, software and IT infrastructure, intellectual property costs, and project-specific expenses. CAPEX includes construction and expansion of fabrication plants (fabs), machinery and equipment purchase, technology upgrades, infrastructure development, facility maintenance, environmental and safety compliance, and new

[107] Similarly, although these numbers are not captured in the figure, Intel's investments were similarly massive around this time. In 2023, the company poured $25.8 billion into CAPEX, equal to about 48 percent of its revenue. And, during this period, it invested $16 billion in R&D—representing nearly 30 percent of its total sales (Intel Corp. 2024).

expansion projects. The Consumer Price Index (CPI) was used to adjust these figures to 2022 dollars.

Sources: TSMC 2023c; BLS. n.d.

Despite the monumental investments represented by Figure 6.1, which made barriers to entry for potential competitors overwhelming, profit margins in the pure-play foundry business were historically cyclical and dependent on node leadership and utilization and it took economies of scale to make TSMC into a commercial juggernaut. This was for several reasons.

First, consider the sector's extreme capital intensity. While the marginal costs of producing additional chips are relatively low once a fab is operational, manufacturing remains highly resource intensive. It demands vast amounts of ultrapure water for rinsing silicon wafers and maintaining cleanroom processes (see TSMC 2023b). It also calls upon a sophisticated supply of rare earth materials that surges by up to 60 percent for advanced nodes (Wiseman et al. 2025). Furthermore, the sector faces a "historic talent deficit," with a projected shortage of tens of thousands of specialized engineers required to operate these complex facilities (ibid).

Second, take the need to keep fabs highly utilized, achieving efficiency gains from large volumes and hedging against financial volatility tied to any single product's market performance (Miller 2022). TSMC's positioning as a contractor that produces a diverse range of chips for various clients—from smartphone chip designers to those inside PCs and GPS units—helped it satisfy these various objectives. To those ends, its foundries were strategically designed to operate for decades, serving a diverse mix of customers by balancing high-end, specialized chip production with long runs of lower-end, commodity chips (Mocker and Mowery 2004; TSMC 2023a).

Lithography Equipment

The microeconomics of the lithography segment of the smartphone chip supply chain are quite different from those associated with their fabrication. Because competition at the leading edge in chipmaking revolves around packing ever more transistors onto chips—requiring patterning at ever smaller feature sizes—this has pushed the industry toward ever shorter-wavelength lithography and the specialized firms capable of building those tools (Miller 2022). Therefore, rather than wafer-throughput economics, this is a high-fixed-cost, low-volume capital-equipment business.

The Dutch firm ASML focuses on producing the extreme ultraviolet (EUV) lithography equipment used for the most advanced chip manufacturing at leading fabs such as TSMC (ibid).[108] EUV systems are colossal, modular machines—on the order of about 180 tons and shipped in many subassemblies—whose cost is well into nine figures per tool, reflecting the extraordinary precision they must deliver at scale (ibid). They use EUV light at a 13.5 nm wavelength to etch circuit patterns onto silicon wafers, enabling the dense, power-efficient transistor geometries that underpin modern smartphone processors (ASML 2025). Achieving this involves an unusually complex process: EUV tools generate light by firing tiny droplets of

[108] Despite its present dominance, ASML's journey has been long and challenging, with decades and billions of euros invested in perfecting EUV. The company delivered its first EUV system in 2010, but it wasn't until 2018 that EUV began large-scale commercial use—a full decade behind initial projections. By 2020, it had delivered only its 100th system. To overcome financial challenges, ASML sold a 23% equity stake to Intel, Samsung Electronics, and Taiwan Semiconductor Manufacturing in 2012, effectively investing its biggest customers in its success. On all these points see Cohen (2025).

molten tin into a vacuum and striking them with high-powered lasers to create a plasma that emits 13.5 nm EUV radiation (ibid). That light is then directed through an ultra-precise all-mirror optical path (because EUV is absorbed by air and most materials), ultimately projecting patterns onto photoresist on the wafer with extreme fidelity (ibid.).

As global demand for advanced processors grew, ASML's technological edge in EUV lithography placed it at the center of the digital economy. Two decades of co-development, a vast supplier base, and an accumulated body of EUV-specific know-how created prohibitive entry barriers—most visibly in the fact that ASML is effectively the sole supplier of EUV lithography systems (Miller 2022). EUV tools incorporate extraordinary complexity (often described as on the order of 100,000 parts/components across subsystems), which helps explain why this segment of the smartphone chip supply chain supports unusually strong margins relative to many other hardware businesses (ASML 2024). Once an EUV platform is developed and in production, the marginal cost of producing additional units is far below the average cost, which can only be recovered through high selling prices and a global customer base. Because ASML commands the market power needed to price well above marginal cost, it has been able to sustain unusually high markups on EUV tools (ibid).

Other Supply Chain Elements

After a clean silicon wafer has been exposed by a lithography tool from a company like ASML—which essentially "writes" the circuit blueprint onto it—it then moves through a series of etching, material-deposition, and cleaning steps (Ayodele 2025; Sharma and Mahajan 2024). These processes are supplied heavily by firms like Lam Research, whose equipment deposits ultra-thin films, etches microscopic features, and cleans wafers between layers to keep yields high as complexity rises (Lam Research 2024). Lam's gross margins have commonly sat in the

mid-40% range in recent years, consistent with a capital equipment niche defined by high R&D intensity and steep barriers to entry (ibid).

Finally, there are players in the smartphone semiconductor supply chain that specialize in advanced packaging. They undertake sophisticated methods for assembling chips that require huge investments. This includes 3D stacking (layering chips on top of each other), using chiplets (smaller, specialized mini chips combined into one larger unit), and interposers (a connecting base that links multiple chips together). Outsourced Semiconductor Assembly and Test (OSAT) companies, such as ASE and Amkor, provide packaging and testing services for semiconductor chips, using techniques like System-in-Package (SiP) and fan-out wafer-level packaging (FOWLP) to improve performance and power efficiency.

The advanced packaging segment offers increasingly attractive profit margins compared to traditional chip packaging. This segment adds significant value by enabling greater chip integration, enhanced performance, and miniaturization beyond what traditional chip scaling alone can achieve. Techniques like 3D stacking and chiplets are complex, require specialized expertise and equipment, and are critical for meeting the demands of modern, high-performance devices.

INTEROPERABILITY HELPED THE DIGITAL ECONOMY EMERGE

While Chapter 7 explores app ecosystems, digital platforms, and the gig economy in detail, I end this chapter with an analysis of an unsung hero in the smartphone saga: the development of the 4G network. That's what ultimately made this device a world transforming technological

breakthrough, setting the stage for the AI era.[109] Ubiquitous, lightning-fast mobile broadband fundamentally altered what could be done on a phone—from streaming video and social media to real-time navigation and cloud services—and allowed it to function smoothly on the go. As mobile data traffic soared, mobile services became an increasingly dominant share of both innovation and the economy (Eisenach and Kulick 2020).

Specifically, 4G LTE, which was launched in 2009 in Stockholm and Oslo, and widely deployed from 2010 to 2015, brought about broadband-class speeds, ranging from 10 to 100 Mbps, and significantly lower latencies, allowing data to travel with much less delay. A crucial underlying change was the shift to "all-IP" (Internet Protocol) architectures—the uniform treatment of all data, voice, and video traffic over the mobile network as internet traffic. In other words, LTE transformed mobile from having a voice channel with a modicum of data attached (the 3G era) to having voice as one application among many: voice was turned into data (see P1 Security 2025). This made the network operate more like the internet users accessed on computers.

---

[109] To be sure, the 3G approach, prevalent from roughly 2001 to 2009, was itself a watershed: the first mobile network to deliver broadband-class data speeds. The third generation of cellular coverage broke the critical 100 kilobits per second (kbps) barrier—which meant data could flow fast enough to enable early smartphones to provide users with access to reliable mobile email and rudimentary app stores. However, while 3G offered speeds that, at their peak, could reach around 2 megabits per second (Mbps) for mobile users, a respectable, but not lightning-fast, broadband connection for its time, these speeds were still relatively modest and, moreover, network coverage remained spotty (Wikipedia 2025a).

For example, achieving reliable round-trip times below 100 milliseconds means that voice-based AI assistants can process your request, send it to the cloud model, and return a spoken reply almost instantly. Likewise, sustaining 5–10 Mbps of throughput ensures uninterrupted high-definition video streaming or real-time AR overlays on maps without buffering pauses, even as users move at highway speeds.

That powerful combination of speed, responsiveness, and internet-centric design engendered a surge of new mobile capabilities: rich, on-device app ecosystems, smooth video streaming, accurate location-aware services like real-time maps, and, most crucially, reliable cloud offload—complex tasks and heavy data processing could be shifted to powerful remote servers without users noticing any interruption. This fundamental shift transformed the smartphone from a device primarily for mobile email and basic apps into an always-connected, AI-enhanced platform.[110]

Indeed, digital platform-based companies like Uber, Instagram, Snapchat, and the countless others I will explore in Chapter 7 could not have scaled without reliable mobile broadband for all users. The app-store model (Apple's App Store and Google Play) only blossomed during the 4G era because consumers came to expect broadband-quality experiences—including Netflix-style video streaming, sub-second load times, and live turn-by-turn navigation. This incentivized developers to build ever more sophisticated, always-on, context-aware experiences.

[110] To be sure, Siri launched on 3G in 2011; but it only truly excelled on 4G; and Google Now followed in 2012.

The advent of 4G also set the stage for the AI revolution. Suddenly, companies like Apple and Google could rely on high-performance, always-on data connections to build sophisticated, AI-powered assistants, a topic I will return to in Chapter 11. And 4G's all-IP, low-latency networks became the essential plumbing for the early AI revolution: fast enough to send voice queries to cloud-based models and receive spoken answers in real time, reliable enough to push continuous model updates to devices, and ubiquitous enough to generate the vast streams of user data—searches, locations, interactions—that would train the next generation of AI systems.

In line with the main thrust of this chapter, this vindicated the trend towards supply chain disintegration. Before 4G, handset makers and network operators still tended to own most of their technology stack end to end—even if, in practice, they were gradually outsourcing bits of ASIC design or middleware by the late 1990s (Funk 2009).[111] Companies like Siemens, Ericsson, Nokia and NEC each sold radios, base stations, handsets and proprietary software platforms under a single corporate roof. To enable interoperability and avoid patent disputes, these companies at first engaged in licensing and cross-licensing arrangements with one another; but as 3G standards proliferated, these one-off agreements became increasingly unwieldy, paving the way for the multilateral approach of the 4G era vested in SSOs and patent pooling (Lévêque and Ménière 2008).[112] I now turn to explaining how this infrastructure came into being and how

[111] ASIC (Application-Specific Integrated Circuit) design refers to the process of creating a custom integrated circuit tailored for a specific application or function, rather than a general-purpose chip like a CPU or GPU.

[112] While patent pools complement SSOs by aggregating essential patents from multiple holders into a single licensing package, allowing implementers to obtain rights through one agreement

the very process of establishing a web of 4G standards reinforced the highly specialized smartphone supply chain that this chapter aims to understand.

By 2025, the mobile wireless industry was composed of various specialized firms, each leveraging their comparative advantage to contribute to the final product and service. This ecosystem included chip and component manufacturers such as Qualcomm, MediaTek, and Broadcom, which designed and produced specialized baseband chips crucial for wireless cellular communications, while other manufacturers focused on specific components incorporated into mobile devices, such as cameras and sensors. Device manufacturers like Apple, Samsung, and HTC integrated these diverse components into complete mobile devices including phones and tablets and developed the user-facing features. Network infrastructure providers built base stations and servers that allowed mobile devices to communicate, with their products incorporating the same core communication components, such as baseband chips found in devices, along with specialized functions that ensured interoperability between systems and maintained quality of service across the network. Finally, network operators such as Verizon and

---

rather than negotiating separately with dozens of patent owners, in the rest of this chapter I will focus on SSOs. However, I should note that although bilateral licensing has dominated handset negotiations—given the small number of major licensors like Qualcomm, Ericsson, and Nokia—patent pools have proved more consequential in newer markets such as automotive telematics and IoT devices, where implementers lack established relationships with cellular SEP holders (Mallinson 2021). The pools' principal contribution is reducing transaction costs and providing royalty predictability by offering a known per-unit fee.

AT&T deployed and maintained wireless networks, managing user subscriptions and interactions to deliver mobile services to consumers.

By staking out industry leadership positions in the SSOs that developed these standards, upstream semiconductor design firms were able to develop business models centered on absurd amounts of R&D and intangible capital. Declaring their pioneering, yet interdependent, inventions as essential to key interoperability standards enhanced the enforceability of their patent portfolio by, among other things, oft-providing readymade injunctive relief.[113] As outlined in Chapter 3, fabless chip designers can lean on US courts to order alleged patent infringers to cease and desist manufacturing copycat goods; similarly, plaintiffs can ask the ITC to block allegedly infringing products from being imported to the United States.

By the same token, these SSOs established FRAND licensing terms that helped solve distributional conflict by setting limits on how much of a stake in the industry profits these upstream designers could claim versus the downstream implementers (see Chapter 3). This delicate balance gave semiconductor idea generators the incentives and opportunities to work in

[113] To be sure, SEP status does not confer a "free" injunction: In practice, FRAND commitments and equity-based doctrines can limit or condition injunctions, particularly where the implementer is deemed a "willing licensee" that engages in good-faith negotiations with an SEP holder—responding to licensing offers and making counteroffers that align with FRAND principles—rather than seeking to delay or avoid payment. Because the doctrinal tests and enforcement norms differ across the United States, the EU, and major Asian venues, SEP leverage often operates less as guaranteed exclusion and more as bargaining power shaped by forum, posture, and willingness-to-license determinations (Nikolić 2017).

concert with other supply chain players to architect the types of interoperability and functionality that made smartphones much more valuable than they otherwise would have been in the absence of SSO-enabled coordination (see Galetovic and Haber 2017).

QUALCOMM, SSOs AND SEPs

As I first mentioned in Chapter 3, 3GPP—among other bodies—is an SSO that helps set global cellular standards.[114]To establish a common language between devices and networks and building on its work on previous generations of wireless standards, 3GPP helped establish the 4G LTE mobile connectivity framework (Holma and Toskala 2012).

While 3GPP manages the technical specifications, the disclosure of SEPs and FRAND undertakings are handled under the IPR policies of its Organizational Partners, such as the European Telecommunications Standards Institute (ETSI). SSOs like ETSI facilitate transparency by maintaining registries of declared patents, but they explicitly remain neutral regarding the specific commercial value or essentiality of those patents, leaving licensing rates to be determined by market participants (ETSI 2024).

[114] While the LTE concept was first proposed by Japan's largest mobile telecommunications company, NTT DoCoMo, in 2004, it was implemented by 3GPP, which includes seven regional standards organizations from Europe, North America, Japan, China, South Korea, and India. Meanwhile, the International Telecommunication Union (ITU), a specialized agency of the United Nations, sets the high-level global vision and performance requirements for each generation of mobile technology, including 4G.

3GPP includes several firms and organizations.[115] "Individual Members" include equipment and device vendors such as Ericsson, Nokia, Huawei, Qualcomm, and Samsung, as their real-world deployment requirements drive the performance, reliability, and interoperability goals of each specification. Major network operators—AT&T, Verizon, China Mobile, Vodafone, and others—also participate as Individual Members, ensuring the standards meet carrier needs. Finally, "Market Representation Partners" encompass trade associations, industry alliances and consortiums (e.g., automotive, public safety, utilities), and national standards bodies or regulators whose spectrum, cybersecurity, and policy expertise feed into the 3GPP process.

The suite of underlying standards agreed to by 3GPP members defined everything from radio interface protocols to core network routing—a universal technical grammar that allowed any compliant device to connect to any compliant network worldwide at broadband speeds. The core network relies on the Evolved Packet Core (EPC), including components such as the Mobility Management Entity (MME) and gateways that route data sessions. Voice over LTE (VoLTE) enabled voice calls to run over the all-IP network as data packets. And on the radio side, LTE relied on OFDMA for downlink and SC-FDMA for uplink, with enhancements like MIMO and carrier aggregation to increase speeds and capacity. By finalizing the transition to an all-IP architecture, these standards transformed mobile phones into internet-connected

[115] Research institutes and certification bodies contribute indirectly via their respective national partners rather than joining 3GPP directly—for example, NIST provides measurement labs and test methodologies. On all the distinctions mentioned in this paragraph see 3GPP (2024).

multimedia powerhouses, providing the foundational infrastructure for the app-based platforms and AI services that would follow (Holma and Toskala 2012).

Qualcomm as a Fulcrum Point

Qualcomm played a pivotal role in developing each of the critical LTE functions. Their integrated modems power not only connectivity, but key aspects of the LTE control plane, ensuring that users maintain session continuity during streaming or browsing.[116] Specifically, Qualcomm's technical breakthroughs enable mobile digital devices to seamlessly attach to the network, manage mobility across cells, and execute smooth handovers without dropped connections (Qualcomm Inc. 2014). In essence, Qualcomm's technologies serve as the invisible foundation upon which every carrier, device maker, and app developer relies to bring LTE devices to life (Gupta 2015: 876-77). This industry leadership built on their earlier successes in CDMA and 3G, solidifying their role in each subsequent generation of cellular networking.

During the 2G era, Qualcomm fundamentally reshaped mobile communications by championing CDMA, a technology that allowed multiple users to share the same radio spectrum simultaneously. As we saw in the last chapter, where earlier cellular standards had assigned distinct time slots or frequencies to each call, CDMA let each handset transmit across the full

[116] The control plane in a mobile network refers to the part of the network architecture responsible for signaling and management functions, rather than carrying user data. Its main tasks include establishing and releasing connections; handling device movement across cells (handover); verifying users and securing communication; and assigning radio and network resources for data sessions. In LTE, the control plane works alongside the user plane, which carries actual user traffic (voice, video and data). See 3GPP (2021a).

bandwidth, distinguished only by unique, device-specific codes. This innovation increased network capacity by leveraging the "voice activity" factor—suppressing transmission during pauses in speech—and enabling universal frequency reuse across every cell, offering capacity gains estimated at up to 20 times that of analog systems (Gilhousen et al. 1991).

Building on that success, Qualcomm helped drive the evolution of cdmaOne into the 3G cdma2000 family through the 3GPP2 standards partnership (Qualcomm 2006).[117] The first phase, "1x," brought voice networks into the digital age with modest data capabilities, while successive "EV-DO" (Evolution-Data Optimized) releases pushed peak download speeds into the low single-digit megabit range. Qualcomm structured cdma2000 to work with existing CDMA deployments, carriers could upgrade without overhauling their spectrum holdings or core networks—a compatibility advantage Qualcomm explicitly touted as making migration "simple and affordable" (ibid).

To fully understand Qualcomm's later influence over LTE, first consider how the 3GPP standardization process works. In a structured multi-stage process, member companies contribute technical proposals that may eventually become part of global mobile standards. The work

[117] While CDMA2000 gained significant traction in North America with operators like Verizon Wireless and Sprint, WCDMA (Wideband Code Division Multiple Access)—the radio access technology behind UMTS (Universal Mobile Telecommunications System)—ultimately achieved far broader global adoption due to its natural evolution path from the globally pervasive 2G GSM (Global System for Mobile Communications) networks, making it the preferred upgrade choice for the majority of operators worldwide, including major European carriers like Vodafone (see Qualcomm 2006).

proceeds from requirements, to overall architecture, and finally to protocols and interfaces that specify message formats, procedures, and timing so devices and network equipment from different manufacturers can interoperate (3GPP 2021b).

At each stage, member companies submit contributions—ranging from detailed text proposals to simulation data and trial reports—that are debated across working groups. Once the final text is approved and locked in via the Change Request process, it generates absolute certainty: the specification becomes a mandatory rule, and every chipset, network element, and device must implement the specified procedures and parameter values to claim compliance (3GPP 2019). This rigidity de-risks the technology, giving major supply-chain players the confidence to invest billions to scale global infrastructure.

By the late 2000s, as 3GPP members gathered to define the next generation of cellular standards, it became evident that the CDMA approaches which had defined the 3G era were no longer optimal for the new mobile-broadband targets. Instead, the LTE standard coalesced around Orthogonal Frequency-Division Multiple Access (OFDMA) for the downlink (Dahlman et al. 2011). Although Qualcomm initially favored a CDMA-based evolution for 4G, it executed a decisive strategic pivot to ensure it did not lose its leadership position. In 2006, the company acquired Flarion Technologies, a developer of "Flash-OFDM," for approximately $600 million to secure a foothold in the new technology stack (Qualcomm 2006). Following the acquisition, Qualcomm poured resources into adapting its legacy innovations to the OFDMA template, developing foundational inventions related to spectral efficiency and the sophisticated control plane functions that optimized OFDMA for mobile environments (ibid).

These investments allowed the company to reassert its dominance at the cellular standards table. Qualcomm deployed a large contingent of highly skilled engineers to 3GPP

meetings, where they presented a high volume of technical contributions and held leadership positions, such as rapporteurs, in key working groups. Qualcomm formally disclosed hundreds of LTE-related patent families to the ETSI database. These declarations covered the critical bottlenecks of the new standard, such as downlink scheduling, advanced MIMO processing, and mobility protocols within the Evolved Packet Core (EPC) that managed how devices handed off connections between towers (see Dahlman et al. 2011).

For the uplink, 3GPP selected SC-FDMA (Single-Carrier Frequency-Division Multiple Access) for its superior power efficiency—an especially binding constraint for battery-powered smartphones (Myung et al. 2006).

By early 2010, Qualcomm held approximately 11 percent of all declared LTE SEPs—and the highest number of "truly essential" patents of any firm in the industry (Cyber Creative Institute 2013). This effectively transposed their monopoly on "essential" IP from the 3G CDMA world into the 4G OFDMA world. And beyond mere volume, the nature of these patent declarations was strategically vital. According to Phelps and Miloni (2012), Qualcomm's LTE portfolio ranked the highest in "essentiality" and the second highest in "novelty" (behind Nokia), particularly for system-level inventions that governed how the entire device interacted with the network—rather than just the internal workings of a modem chip.

By securing patents on the system rather than just the component, Qualcomm laid the groundwork for its licensing model. Because their IP covered the device's fundamental ability to communicate, they could justify calculating royalties based on the wholesale price of the entire phone, rather than the significantly lower cost of the chipset itself.

4G IP and Licensing Regime

During 4G negotiations, a consortium of five major telecom firms—Ericsson, Nokia, Alcatel-Lucent, NEC, and NextWave Wireless—proposed a framework to cap aggregate LTE royalties at a "single-digit percentage" of the handset's wholesale cost (Alcatel-Lucent et al. 2008). This proposal was particularly attractive to "net licensee" companies—device manufacturers that pay out more in royalties than they receive from their own patent portfolios (Stasik 2010). It was vehemently opposed, however, by patent holders like Qualcomm; it argued that not all patents contributing to a standard are of equal value and licensing terms should be determined by bilateral negotiations rather than arbitrary industry-wide caps (Qualcomm 2008).

Ultimately, LTE commercially launched without an industry-wide, formally agreed-upon royalty cap. Qualcomm ultimately settled on a two-tier licensing strategy—a cellular SEP-only license at 3.25% of the net selling price of handsets and a full portfolio license at 5%, with all programs including a per-device cap (Qualcomm Inc. 2024).

However, Qualcomm's "patent licensing business model actually sparks competition by allowing all kinds of companies to easily enter the smartphone business without having to make big investments in core cellular research and development" (Freeman 2018).[118] Moreover, Owen (2025) argues that standards in general establish common interfaces that allow components from different manufacturers to work together seamlessly; this modularity ensures that downstream firms can switch partners or suppliers without needing to redesign their products entirely, reducing the risk of being locked into a specific relationship and reducing barriers to entry. A relatively wide pool of contributors to the LTE standard meant that any one firm would be hard pressed to dictate licensing terms. According to Cyber Creative Institute (2013), even

[118] Owen 2025 formalizes this logic.

Qualcomm, while clearly the market leader, only held between 11 and 15% of truly essential LTE patents, followed by a diverse field of competitors including Huawei (about 9%), ZTE (about 8%), Nokia (about 8%), and others like InterDigital (about 7%).

Nonetheless, every 3GPP member who declared a patent essential to LTE had to formally pledge to license that patent on FRAND terms without specifying what "fair" or "reasonable" meant (ETSI 2019). Therefore, individual patent holders like Qualcomm structured their own licensing agreements to be "reasonable" by including per-device caps. Historically, the price for a comprehensive Qualcomm license was typically 5 percent of the wholesale price of the device, subject to a cap, which effectively limited the maximum royalty to $25 per phone (based on a $500 device cap) (Stasik 2010). However, starting in late 2017, the company began offering two distinct licenses: one for its SEPs at a rate of 3.25 percent of the wholesale device price, and a second for its entire portfolio at the original 5 percent (Qualcomm 2017). Concurrently, Qualcomm lowered the per-device cap from $500 to $400 (Qualcomm Inc. 2019a). This change meant that, circa 2025, an SEP-only license peaked at $13 per phone (3.25% of $400).

Qualcomm's revenue model reflected this principle. Circa 2025, the company typically licensed its extensive patent portfolio as a comprehensive package to handset manufacturers, charging royalties as a percentage of the wholesale price of each device, irrespective of whether that specific device used a Qualcomm chip (Sullivan & Cromwell LLP 2020). The device maker was obligated to compensate Qualcomm for the right to utilize its patented technologies, which were embedded not only within baseband chips themselves, but also throughout other components of the phone or cellular device (Gervase et al. 2020). The legality of Qualcomm's

licensing model, particularly its "no license, no chips" policy and device-level royalty structure, was largely vindicated by a 2020 circuit court ruling (see Qualcomm 2020).[119]

THE HOLDUP PROBLEM AND THE ROLE OF STANDARDS

By possibly functioning as commitment devices that alter payoff structures, technical standards like 4G-LTE offer a potential solution to credibility problems that arise within supply chains where opportunism is endemic. As I outlined in Chapter 3, a firm operating within a fractionalized supply chain may fear that, once they've incurred fixed, irreversible, and asset-specific investments, other players may threaten their quasi-rents—the wedge between a firm's short run average costs and long run average costs that arise when investments are sunk and cannot be redeployed to alternative uses (see Galetovic and Haber 2017, who build on Klein et al. 1978 and Williamson 1985). Because firms split the producer surplus that they jointly create in the act of bringing new products to market by accepting binding shared rules—instead of letting the market purely decide or by engaging in unregulated bargaining—their cooperation is rendered more credible: standards may effectively stretch out firms' time horizons, enhancing the value they place on future revenue streams, eclipsing the short run gains from holding their partners up.

[119] As discussed in Chapter 3, this decision vacated an earlier ruling favoring an FTC antitrust lawsuit, affirming that Qualcomm was not obligated to license rival chip suppliers, its practices were "chip-supplier neutral" and did not unfairly surcharge rivals' modem sales, and that prior agreements with Apple, which had been a party to the suit, had not substantially foreclosed competition.

While the SEPs declared by upstream semiconductor firms during SSO proceedings may help them secure returns to their investments, these institutional safeguards remain incomplete.[120] As Chapter 3 made clear, the massive sunk costs in developing new mobile chip architectures and wireless technologies create significant quasi-rents that downstream firms might successfully appropriate despite ostensibly strong IP rights. Qualcomm typically invests five to seven years in advance of a new product release and spends billions of dollars developing each new generation of mobile technology before seeing any returns, with $8.2 billion spent in 2022 alone, representing nearly 20 percent of its revenue (Qualcomm Inc. 2022b). While its licensing model aims to recoup these massive upfront investments, downstream firms such as Apple may weaponize litigation to nibble away at its margins.

What explains this asymmetry? While Qualcomm's R&D investments are specialized and sunk, downstream device assemblers may face relatively low switching costs given the modular

---

[120] This claim inverts the orthodox "patent holdup" thesis, which holds that upstream SEP holders exploit downstream manufacturers by demanding excessive royalties after implementers have made standard-specific investments (see Shapiro 2010). Galetovic and Haber (2017) demonstrate that this orthodox view conflates two mutually inconsistent economic mechanisms—holdup (appropriation of quasi-rents) and the exercise of market power (extraction of monopoly rents)—and rests on logical fallacies, including the elision of opportunistic surprise as a necessary condition for holdup. They further show that the theory's core predictions are falsified by empirical evidence: SEP-intensive industries exhibit rapid innovation, falling quality-adjusted prices, robust firm entry, and cumulative royalty yields far below what patent-holdup models predict.

ARM architecture, which licenses its instruction set to multiple chipmakers, allowing them to design compatible chips while maintaining standardization across the industry. Qualcomm's quasi-rents therefore remain vulnerable to appropriation by downstream implementers because litigation is costly, FRAND terms can be disputed ex post, and SEP enforcement is variable and potentially politicized.

Perversely, downstream firms often accuse upstream innovators of the very holdup that they themselves engage in (see Galetovic and Haber 2017; Galetovic et al. 2024). For instance, between 2017 and 2019, Apple and Qualcomm engaged in a global legal battle over patent licensing fees, with Apple claiming Qualcomm's royalty rates were excessive in several high-profile lawsuits (see Qualcomm Inc. 2020). The dispute ended with Apple paying Qualcomm approximately $4.5 billion and agreeing to a six-year patent licensing agreement (Qualcomm Inc. 2019b), but not before the latter incurred millions of dollars in legal fees. Moreover, the challenges that upstream firms face when attempting to secure a return on their investments are aggravated by intense lobbying from downstream assemblers seeking to weaken their IPRs; for example, by efforts to weaken injunctive relief or impose limits on royalty rates (see ACT 2021; Barnett 2017).

To mitigate these risks, upstream firms deploy strategies beyond participating in SSOs. Qualcomm exemplifies this approach: it diversified its customer base across multiple manufacturers (Samsung, LG, Motorola) to reduce dependence on any single buyer; developed complementary software tools and developer ecosystems (such as BREW and system reference

designs) that raised switching costs;[121] entered cross-licensing agreements that aligned incentives by intertwining patent portfolios (Qualcomm Inc. 2018; Layne-Farrar et al. 2007); maintained a reputation for sustained, well-resourced litigation that signaled opportunism would be costly (AIPLA 2019; PatentTrack 2024);[122] and invested relentlessly in R&D to stay ahead of the technological curve. The 2019 Apple-Qualcomm settlement—which saw Apple pay approximately $4.5 billion and agree to a six-year licensing arrangement after extensive

[121] BREW (Binary Runtime Environment for Wireless) was a mobile application development platform created by Qualcomm and launched in 2001. It provided a standardized software environment that allowed third-party developers to create applications for CDMA-based mobile phones across different handset manufacturers (see Qualcomm Inc. 2006). BREW represented an early precursor to modern app ecosystems like Apple's App Store and Google Play, though it operated through a more curated, carrier-controlled distribution model.

[122] Qualcomm's litigation strategy has not always been successful or admired. As an example, consider that a lawsuit it filed against Broadcom Corp. in a patent infringement allegation devolved into a discovery dispute that led to findings that Qualcomm had withheld over 300,000 pages of responsive documents. The case spanned several years and resulted in an $8.5 million sanctions order against the company, later partially vacated (see Benesch et al. 2008). The episode nonetheless illustrates how extensive, document-heavy, and costly Qualcomm's standards-related patent disputes can be.

litigation—illustrates both the vulnerability of upstream quasi-rents and the ultimate efficacy of these defensive strategies (Qualcomm Inc. 2019b; Qualcomm Inc. 2019c; Nellis 2023).[123]

CONCLUSION

As we saw in Chapter 4, the CDP era was characterized not by the monolithic, vertically integrated giants of previous eras, but by finely disintegrated global ecosystems in which each participant played a specialized role. Advances in information and communications technologies, liberalized trade and capital flows, and robust international IP regimes created economic incentives and legal safeguards for upstream innovators to concentrate on intangible assets, including architecture design, proprietary algorithms, and patent portfolios—while outsourcing the capital-intensive tasks of fabrication, assembly, and distribution. In turn, deep specialization unleashed innovation at a scale and pace far beyond that achievable by any single vertically integrated entity.

As this chapter argues, the smartphone supply chain circa 2025 embodies the intersection of both revolutionary currents. Moore's Law-related miniaturization efforts and their associated economic imperatives—especially soaring R&D costs and capital investment requirements—accelerated product cycles while shortening the commercial lifespan of specific chip generations. The central problem was therefore one of risk: it became difficult for chipmakers to predict whether demand for a given product would be high enough, and last long enough, to keep specialized production capacity fully utilized long enough to recoup R&D expenses and fixed

[123] To be sure, Apple acquired Intel's smartphone modem business to develop in-house capabilities (Apple 2019b). However, demonstrating the relative effectiveness of Qualcomm's protean IP strategy.

costs. This amplified the financial peril of committing to bespoke technologies and facilities and drove sharper functional separation across the value chain. Distinct stages of chip production evolved unique cost structures, investment requirements, and strategic imperatives: while design commanded outsized returns on sometimes relatively modest R&D investments, fabrication and equipment required staggering sunk costs and yielded only thin margins, making it clear that no single firm could—or should—own every link in the chain.

In this context, and riding on the stronger IP and globalization pushes outlined in Chapter 3, the smartphone's development separated design, component manufacturing, software development, assembly, and application. The advent of advanced software made it cheaper and easier to design chips without having to manufacture them, and strong demand for smartphones—along with generous subsidies furnished by the Taiwanese government—allowed TSMC to achieve economies of scale around high-volume contract production. Companies like Apple focus on design and marketing while outsourcing assembly to contract manufacturers such as Foxconn and Pegatron. Specialized firms handle component production—Corning supplies Gorilla Glass, Qualcomm designs processors and modems, Sony manufactures camera sensors, and TSMC fabricates custom-designed chips. Fabless chip designers like Qualcomm and NVIDIA rely on Electronic Design Automation firms such as Synopsys and Cadence to create complex chip designs, which are then manufactured by pure-play foundries like TSMC and Samsung.

This chapter also documents how interoperability standards supplied the coordination layer that allowed this fragmented ecosystem to function globally. SSOs did not "guarantee" seamless interoperability on their own; rather, they set technical specifications and compliance expectations that firms implemented, enabling networks, handsets, and applications to work

across borders and across rival ecosystems—not to mention that a call could go through even if it was made between an iPhone and a Samsung Galaxy.

By the 2010s the result was a revolutionary device made possible by disaggregation, specialization, and the coordination of thousands of firms across dozens of political jurisdictions that sold to hundreds of millions each year for almost two decades, transforming the world in the process. While Apple, the industry pioneer, tightly controlled the hardware and software experience through a closed, proprietary, and highly curated user interface strategy, it nevertheless depended on open connectivity standards and protocols beneath the surface, with the most prominent being those governing cellular coverage, as well as outsourced components that brought those standards to life by fabless chip designers like Qualcomm.

Indeed, while I will examine Apple's business model in more depth in Chapter 8, explaining how end-to-end control improved reliability, security, and monetization of the developer ecosystem, this chapter reveals the company's deeper competitive edge: orchestrating multi-country, high-precision supply chains at scale. This not only freed the firm to specialize in idea-intensive functions—hardware and software design and cointegration, ecosystem building and governance, brand building, and customer experience—but enabled it to harness the massive capital investments and specialized R&D of its partners, thereby delivering cutting-edge hardware at a global scale and cost-efficiency that no single, vertically integrated firm could have achieved alone.

This feat was itself a key building block of the Fourth Industrial Revolution because it made premium hardware widely available and cost efficient, connecting billions via cellular networks and locating users with GPS. Mobile, always-connected, GPS-guided supercomputers—equipped with cameras, location services, and app ecosystems—generated

continuous streams of data about communication, movement, consumption, and interaction that could be aggregated and analyzed at scale using machine learning. This complemented parallel advances in LLMs, cloud computing and other pivotal elements of the AI supply chain, which I will explore in Chapter 8.

CHAPTER 7: AI's ANTITRUST SEEDBED

As I briefly laid out in Chapter 3, a platform brings together different types of users (or market sides), such as consumers and merchants. It enjoys direct network effects when the members of one of those groups are better off as a function of other members of the same group joining in greater numbers—for example, how a payment network becomes more useful to a cardholder as more of their peers adopt the same card, facilitating easier peer-to-peer transfers or shared benefits. Conversely, a platform enjoys indirect network effects when it becomes more valuable to the members of one group because the members of another group join in large numbers. Continuing with the credit card example, as the number of cardholders grows, it attracts a greater number of merchants seeking to capture their spending; as more merchants accept the card, it becomes even more indispensable to consumers (see Rysman 2009).[124]

---

[124] For any multi-sided platform, the initial and most significant hurdle is the "chicken-and-egg problem": consumers won't join a platform with no sellers or content, and sellers won't join a platform with no consumers. To solve this, platforms must adopt a deliberate launch strategy to attract a "critical mass" on one side of the market first. One common approach is to attract the "producer" side of the market; for example, a new shopping mall will sign leases with reputable anchor stores to guarantee a supply of goods that will then draw in shoppers. Another strategy is to subsidize one group of users, such as a video game console maker selling its hardware at a loss to build a large player base, which in turn attracts game developers who pay royalties to access that market. The same network effects that fueled the platform's rise can work in reverse, however, creating a death spiral where the departure of members of one side of the market may drive away those on other sides, causing the platform to collapse. Therefore, attracting users is

During the first part of the 21st century, the rise of world-spanning digital platforms set off a Cambrian-like explosion of data that paved the way for the Fourth Industrial Revolution. Algorithms personalize users' experience through recommendation systems and targeted advertising, making the ecosystem so convenient and tailored to the user that leaving it would mean a significant loss of data and functionality. In turn, a highly engaged user base, and the vast amount of data it generates, becomes the platform's most highly valued asset, and is precisely what advertisers and third-party developers need to succeed. Advertisers gain access to rich consumer data to personalize their ads and test their effectiveness, while developers use the data to build more desirable and commercializable apps. This creates a powerful flywheel that I will take up in Chapter 8: integrated ecosystems lock in users and generate insightful personal data, which in turn attracts the advertisers and developers who make the ecosystem even more valuable, reinforcing the platform's dominant market position. The ultimate knock-on effect is a readymade supply of data that helps train AI models.

Consider that a fundamental challenge facing LLMs is their need for extensive, diverse data to recognize grammatical structures, semantic relationships, and contextual nuances. As the next few chapters will argue in depth, real-time data flows enable LLMs to continuously learn and adapt, improving their ability to predict user behavior patterns. While static datasets like

---

only the first challenge; keeping them connected also matters, as well as balancing each side by maintaining the right ratio of different participants. While regulation can help platforms emerge, grow, and endure, deliberate measures to improve the user experience are also necessary to keep platforms thriving—a topic I will explore in the next chapter.

operating manuals and books certainly help AI models achieve better generalization and precision, learn from edge cases, and reduce bias, real-time behavioral data—including social media posts, search history, and purchase patterns—significantly amplify their ability to do so, allowing models to generate increasingly human-like text and responses (Ouyang et al. 2022; Baek et al. 2024; Salemi et al. 2024). By processing live, dynamic interactions, geo-coded digital platforms detect behavioral shifts, deliver timely recommendations, and enhance personalization, creating more contextually relevant user experiences and, thus, generating "better" data.

While the previous chapter focused on the digital leviathan's physical "skeleton"—the specialized hardware and global supply chains that gave birth to smartphones—this chapter takes up the critical shifts in antitrust litigation and enforcement that represent the "legal oxygen" allowing platforms to inhale users and exhale data. By replacing a reflexive hostility toward large firms and creative business strategies with a "consumer welfare" standard that also rewarded scale and innovation, the US legal system permitted these platforms to achieve the massive network effects necessary to become the primary data-vacuuming engines of the AI era, a topic I will revisit in the next chapter. Eventually, antitrust practitioners grew to tolerate, and in some cases even advocate for, the market dominance of high-technology firms. In doing so, they were inspired not only by Schumpeterian economics, but by strong evidence for efficiencies associated with both platforms' acquisitions of competitors, suppliers, and distributors to achieve demand side economies of scale and their various experiments creating unique ecosystems.

While Chapter 8 explains how the combination of data volume, variety, and velocity facilitated by digital platforms proved essential for developing sophisticated AI systems, this chapter explores the transition from a populist-statist obsession about firm size and market structure to a technocratic focus on consumer welfare, efficiency, and innovation. Antitrust

decision-makers repeatedly looked the other way when Big Tech firms sought to create elaborate platform ecosystems in the name of growing the pie via aggressive pricing, bundling disparate services, and merging with rivals. High levels of market concentration were not a byproduct of the technology's affinity with network effects alone, but the endogenous result of a century-long antitrust evolution that saw increased size, scope, and integration as virtues to be cultivated. In this chapter I therefore return to discussing the bipartisan, evidence-based policy agenda known as the Creative Destruction Paradigm. As I argue throughout this book, the CDP underpinned deep transformations in global trade, IP, telecommunications, electronic contracting, data governance, and business strategy, with the eventual result being a digital economy centered around interoperable mobile devices, global platforms, and the transmission of billions of terabytes of data taking place at blazing speeds (Statista 2024a).

Take the National Commission for the Review of Antitrust Laws and Procedures (NCRALP) as an example. Established by an Executive Order under President Jimmy Carter, which was tasked with increasing the "speed and efficiency" of antitrust litigation (see Carter 1977). Notably, its membership and staff were dominated not by Chicago School ideologues, but by mainstream and "Harvard School" technocrats, including then-Assistant Attorney General John Shenefield and future Supreme Court Justice Stephen Breyer (see NCRALP 1979). The report's championing of economic efficiency and consumer welfare over populist "small is beautiful" ideals represents a bipartisan technocratic shift centered on cost-benefit analysis and the empirical costs of litigation.

Or consider the 1995 Intellectual Property Guidelines, which introduced the concept of "innovation markets" for R&D and other activities that lead to new or improved products and services (DoJ and FTC 1995). This document and its resulting regulations further entrenched the

idea that preserving innovation was essential to protecting consumer welfare and declared that traditional market analysis, which focuses on current competition and market shares, might overlook the potential harm to future innovation from anticompetitive mergers and practices (see Gilbert and Sunshine 1995).

As I will show further ahead, efforts like these on the enforcement side of competition policy complemented a multi-decade long judicial reckoning: US courts arrived at the conclusion that the mere existence of a monopoly is not enough to trigger the antitrust whirlwind; there must also be evidence of "willful" exclusionary conduct that harms the competitive process itself. In fact, a series of high-profile cases condoned market dominance as a legitimate prize for business success if it was achieved through superior efficiency or innovation.

Explaining the Endogenous Rise of Big Tech

The conventional wisdom is that the Athena of maximizing consumer welfare sprouted immaculately from the brow of the Chicago School's Zeus (see Bork 1978; for a critical perspective on this "immaculate" origin story, see Wu 2018). This chapter instead argues that this antitrust revolution was more a natural evolution that saw important antitrust players increasingly gravitate towards language in the Sherman and Clayton Acts that market power alone is insufficient to outlaw certain business practices or block certain mergers. By extension, the mere presence of market power without evidence of exclusionary conduct suggested to antitrust decision-makers that potential competitors were free to enter the market and compete for the rents enjoyed by a dominant firm. And even if this did not happen, it became eminently conceivable to them that the threat of entry might discipline incumbent firms to price competitively, even in highly concentrated markets (see Baumol et al. 1982).

In this vein, ahead I argue that it was ultimately necessity, not ideology, that led antitrust decision-makers to seek out the logic and language of economics and to settle on a consumer welfare centric approach. Initially, they turned to static price theory to provide concrete answers to real-world cases involving raw materials producers, consumer durables, or sports leagues. However, as industries and business practices grew more complex, they increasingly sought out more sophisticated paradigms. For example, as traditional industries like grocery retailing were disrupted by the rise of chain supermarkets, Industrial Organization theorists began to emphasize the potential benefits of vertical restraints and vertical integration. Finally, the Schumpeterian recognition that disruptive technological change may require firms to chase after outsized profits overlapped with antitrust decision-makers' growing awareness that it was impossible for them to keep pace with rapidly evolving markets, feeding a more conservative, wait-and-see approach. The emergence of the digital platforms and ecosystems that propelled the Fourth Industrial Revolution occurred in the shadows of this silent revolution.

This chapter continues as follows. First, I review some antitrust basics. Second, I trace the 'populist-statist' approach that dominated the early 20th century, characterized by a reflexive hostility toward firm size. Third, I recount the organic transition toward the "Rule of Reason" and the consumer welfare standard, showing how economic necessity and new industrial theories replaced structuralist rigidities. Fourth, I examine the specific judicial precedents and agency forbearances that allowed digital platforms to achieve the massive scale and network effects that define today's economy. I conclude by explaining how this permissive legal environment created the "data gushers" essential for the Fourth Industrial Revolution, while signaling the modern return of populist challenges explored later in the book.

ANTITRUST BASICS AND DIFFERENT APPROACHES

American antitrust laws evolved over more than a century to promote competition and prevent monopolistic practices, with four key pieces of legislation forming the backbone: the Sherman Act, the Clayton Act, the FTC Act, and the Robinson-Patman Act (FTC n.d.a; US Congress 1890; US Congress 1914a; US Congress 1914b; US Congress 1936). Enforcement of these antitrust laws is primarily carried out by the DoJ and the FTC (DoJ n.d.a; DoJ n.d.b; FTC n.d.a). Courts play a crucial role in interpreting antitrust laws and setting precedents that shape future enforcement (see FTC n.d.a).

Legislative Foundations

The Sherman Act, enacted in 1890, is the first and most foundational antitrust law (US Congress 1890). While it prohibits monopolistic behavior and practices that restrain trade, its scope and implications for market structures and firms' behaviors evolved over time—one of the topics I will explore in the remainder of this chapter. It consists of two main sections (see US Congress 1890). Section 1 outlaws "every contract, combination... or conspiracy, in restraint of trade or commerce among the several States," addressing concerted actions that restrict competition (ibid). Section 2 prohibits monopolization, attempted monopolization, or conspiracies to monopolize any part of interstate commerce, targeting single-firm conduct aimed at establishing or maintaining monopoly power (ibid).

The Clayton Act, enacted in 1914, was designed to address Sherman Act limitations; it focuses on specific practices that could lead to anticompetitive outcomes (US Congress 1914a). Key provisions include Section 2, which prohibits price discrimination between different purchasers if it lessens competition or creates a monopoly (US Congress 1914a; US Congress 1936). Section 3 outlaws exclusive dealing agreements and tying arrangements that could significantly lessen competition (US Congress 1914a; FTC n.d.b). Section 7 prohibits mergers

and acquisitions where the effect may be to substantially lessen competition or create a monopoly (US Congress 1914a; FTC n.d.b). Section 8 restricts interlocking directorates, where the same individuals serve on the boards of competing companies (US Congress 1914a; FTC n.d.b).

As an overhaul of Section 2 of the Clayton Act, the Robinson-Patman Act of 1936 aims to prevent anticompetitive practices in pricing and protect small businesses from unfair competition by larger firms (US Congress 1936). It does this most specifically by restricting price discrimination: prohibiting sellers from charging different prices to different buyers if it harms competition by favoring large purchasers over smaller ones (ibid). Of all the pieces of legislation that underpin antitrust enforcement in the US, this one had become the least salient, least enforced laws due to the fact that courts and enforcement agencies eventually viewed its protection of small retailers as a form of "price floor" that perversely harmed consumers by preventing the efficiencies and lower prices generated by large-scale chain stores and wholesalers (see Hovenkamp 2005; FTC n.d.a).

Antitrust Enforcement

Enforcement of US antitrust laws is primarily carried out by the DoJ and the FTC (DoJ n.d.a; DoJ n.d.b; FTC n.d.a). The Antitrust Division of the DoJ is particularly focused on criminal enforcement and civil cases involving monopolistic practices and mergers. It has the authority to prosecute criminal violations of antitrust laws and bring civil enforcement actions in federal court.

In contrast, the FTC enforces antitrust laws primarily through civil enforcement actions, concentrating on preventing unfair competition and protecting consumers (US Congress

1914b).[125] The FTC can bring cases in its administrative court, as well as in federal court, and has broad authority to investigate and remedy unfair business practices through its rule-making and adjudicative powers (FTC 2025). Over time, both agencies tended to converge towards similar norms and strategies such as the prioritization of the Consumer Welfare Standard, which I will touch on further below, and a reliance on sophisticated econometric modeling to predict market outcomes, although not always uniformly or without conflict (see Hovenkamp 2005; Kovacic 2003).

Importance of Litigation to Antitrust

Antitrust enforcement in the US heavily relies on litigation (FTC n.d.a). Lawsuits against firms can be brought by the FTC, DoJ, state attorney generals, or private parties (DoJ n.d.a; DoJ n.d.b; US Congress 1976a). Under the Clayton Act, individuals and businesses harmed by antitrust violations have the right to sue for treble damages; this serves as a strong financial incentive for private actors to help enforce these laws (US Congress 1914a).

Therefore, federal courts are central to antitrust enforcement: they interpret and apply antitrust statutes, thus setting legal precedents that guide future cases and occasionally influence

[125] The FTC Act, also enacted in 1914, established the FTC as an independent executive agency with broad powers to prevent unfair methods of competition and unfair or deceptive acts affecting commerce (US Congress 1914b; FTC n.d.a). The FTC Act's most important provision, Section 5, declares that "unfair methods of competition" are unlawful, as are "unfair or deceptive acts or practices" (US Congress 1914b). This endows the FTC with broad authority to address various anticompetitive behaviors beyond those explicitly covered by the Sherman and Clayton Acts (US Congress 1914b; FTC n.d.a).

Congress to amend these laws in response to judicial interpretations (FTC n.d.a). While lower federal courts handle the bulk of antitrust litigation, significant cases often progress through the appeals process and may ultimately reach the Supreme Court (FTC n.d.a). Seminal Supreme Court decisions about antitrust, several of which I shall examine below, have set binding precedents for lower courts and have significantly influenced antitrust policy, market structures, and firm strategies (see FTC n.d.a).

Across all levels, courts inform their application of antitrust law with various types of economic analysis and empirical evidence, often calling on economists as expert witnesses to assess the legality of business practices (Froeb 2003). Consequently, the courts' interpretation of antitrust laws has served to either broaden or narrow the scope of permissible business conduct and indirectly fostered the emergence of entire new industries (FTC n.d.a), as demonstrated by the rise of digital platforms, one of the main topics covered in this chapter. This judicial influence is further mediated by agency-issued guidelines—such as the "Horizontal Merger Guidelines"; which, while not legally binding, provide the economic frameworks that courts often use to determine if a market structure is likely to harm consumers (ibid).

To understand how this "regulatory greenhouse" for platforms was built, we must first look back at the "populist-statist" tradition that the Creative Destruction Paradigm eventually dismantled.

The Populist Approach to Antitrust

Before examining this approach in detail, it is helpful to recall the core tenet of the populist-statist approach detailed in Chapter 3. This political and policy paradigm was rooted in the belief that the state must actively manage market structures to preserve social and political stability. It prioritized constituency protection—shielding small businesses and local merchants from the

"disruptive" productivity of large-scale industrialization. Concentrated economic power was viewed not as a sign of efficiency, but as a "despotic" threat to the democratic process. By utilizing entry barriers and price controls to engineer a predictable, if static, marketplace, the populist-statist project sought to avoid the very "creative destruction" that the modern era took for granted. It was this atmosphere of cautious *small-c conservatism* that Justice Louis Brandeis and his successors transformed antitrust into a moral crusade against "bigness" in which the government did not just police competition, but actively managed market outcomes to protect the "social" fabric from large-scale industrialization's disruptions.

The Brandeis School of Antitrust

Supreme Court Justice Louis Brandeis, serving from 1916 to 1939, was a key figure in interpreting antitrust laws broadly to curtail the allegedly pernicious effects of big business on not only the economic wellbeing of everyday merchants, workers, and customers, but democracy itself. Brandeis feared that wealthy individuals and families could use their economic clout to manipulate the political system, sometimes through outright bribery, to build trusts that could corner different markets. Infamous examples include J.P. Morgan's railway trust, the Northern Securities Company, and his steel trust, US Steel. Brandeis sought to break companies like this up to put an end to "industrial feudalism" (see Urofsky 2009).

In the 1930s, populists inspired by Brandeis opposed "centralized wealth, centralized control, and centralized location," favoring local self-sufficiency and rural communities over economic specialization and large business firms (Hofstadter 1955: 210). Senator Sherman and Representative Taylor warned against the dangers of large combinations of wealth and power, likening them to autocratic rule, and railed against the perceived ability of big firms to manipulate the political system and exacerbate inequality. Theodore Roosevelt's administration

aggressively enforced antitrust laws, targeting railroad, oil, and banking trusts. His fifth cousin, Franklin Delano Roosevelt, continued this legacy by enshrining trust-busting policies into his New Deal. FDR ran aggressively against big business; in his famous 1936 convention acceptance speech, "A Rendezvous with Destiny," he argued that modern life had been pressed into the service of "economic royalists" who created a "new despotism" (Roosevelt 1936: 32).

Economic thinking at the time also reflected these concerns. Edward Chamberlin's "Theory of Monopolistic Competition" argued that firms competed by excessively differentiating their products, leading to suboptimal outcomes where prices exceeded marginal costs and firms continuously churned out "excess capacity" (Chamberlin 1927). His analysis thus reinforced the populist argument that unchecked corporate power required robust antitrust enforcement to serve the public interest, not only to protect consumers but also to preserve fair competition and democratic integrity (Hawley 1966).

After a lull in going after big and powerful firms during World War II in line with a war effort aimed at rapidly churning out weapons on a massive scale, the postwar era witnessed a resurgence of aggressive antitrust enforcement. A rapid surge in consumer price inflation and concomitant complaints made to the federal government inspired new legislative measures and stricter enforcement to address perceived threats to competition (Kovacic and Shapiro 2000). Between 1938 and 1939, the number of antitrust lawyers at the DoJ's Antitrust Division nearly tripled, and enforcement actions increased significantly (Arnold 1941). The Celler-Kefauver Act of 1950, which aimed to curb "anticompetitive mergers," reflected an ongoing obsession with fears about unchecked economic power (US Congress 1950; Kintner 1978).

Judicial Philosophy and Practice During Populist-statist Era

New concerns about high prices clashed with the government's enduring attempts to shield small, local competitors and their associated political constituencies from the overwhelming productivity of larger, more integrated firms (see McCraw 1984). Therefore, the antitrust jurisprudence of the postwar era was marred by inconsistency and a lack of coherent economic analysis, resulting in a highly politicized enforcement regime (Hovenkamp 2005). Court decisions were driven more by reflexive populist sentiments and a distrust of big business than a rigorous assessment of competitive effects (Posner 1979). Judges categorically and peremptorily forbade business practices such as horizontal information-sharing agreements, vertical resale price maintenance, exclusive dealing contracts, tying arrangements, price discrimination, group boycotts, exclusive territories, and patent pooling (see Bork 1978; Muris 2023). They also sought to socially engineer market structures by dictating the number of firms allowed to compete, shrinking large firms, and protecting small businesses (see Kovacic 1990; Hovenkamp 2005). In doing so, these judges prioritized the static welfare of existing market participants over the dynamic efficiency that fuels long-run innovation.

Consider a few famous cases. In *Alcoa* (1945), the Second Circuit (sitting as a court of last resort) condemned Alcoa for monopolizing the aluminum market, despite strong evidence that the company's dominance resulted from superior efficiency (Crandall and Winston 2003: 14). This decision was belatedly criticized by antitrust scholars, particularly those associated with the Chicago School, as effectively punishing business success rather than focusing on genuinely anticompetitive conduct (Bork 1978; Posner 2001). Similarly, in *Standard Oil* (1949), the court condemned exclusive dealing contracts between Standard Oil and independent gas stations, despite the lack of clear evidence of harm to competition (see Bork 1978; Posner 1976). And in *American Can* (1949), the DoJ accused American Can Company of monopolizing the can

manufacturing industry by acquiring numerous smaller competitors and engaging in exclusionary practices such as exclusive dealing arrangements and restrictive contracts with suppliers. The court agreed and, aiming to explicitly reduce industry concentration and lower barriers to entry to "maintain competitive market structures," mandated that the firm divest itself of certain assets (see Hovenkamp 2009: 273).

Merger law perhaps provides the most striking example of post-war populism. In *Brown Shoe* (1962), the Supreme Court blocked a merger between two relatively small shoe companies with a low combined market share to protect small businesses, regardless of broader competitive effects and consumer benefits. While *Philadelphia National Bank* (1963) was a bit more circumspect, with the court ruling that only mergers resulting in a firm holding a significant market share would be presumed illegal, it still perpetuated the view that certain business decisions were simply verboten. Accordingly, to prevent further market concentration, in *Procter & Gamble* (1967) the Supreme Court prevented Proctor & Gamble from acquiring Clorox, disregarding potential efficiencies and benefits to consumers. Fittingly, dissenting in *Von's Grocery* (1966) after the Supreme Court blocked the merger of two grocery stores operating solely in Los Angeles, Justice Stewart famously remarked that "the sole consistency that I can find is that in litigation under Section 7 [of the Clayton Act], the Government always wins" (see Von's Grocery Co. 1966: 301).

These cases underscore that, during the post-World War II era, the structural approach to industrial organization became a central tenet of the Populist Antitrust tradition. Economists such as Joe Bain (1950; 1956; 1959) and E.S. Mason (1964) played pivotal roles in shaping this perspective. They identified various barriers to entry, including economies of scale and incumbent advantages like superior technology and product differentiation, as key determinants

of market power. Bain and Mason then posited a linear relationship between market concentration and economic outcomes: when the number of sellers decreased, output fell, and prices rose. This structuralist view also held that high market concentration indicated significant barriers to entry, which led to excessive product differentiation and a reluctance to compete on price. The result, they argued, was a market characterized by high short-run profits, overinvestment in advertising and branding, and overall stagnation (see Hovenkamp 2009).

Donald Turner, who led the Antitrust Division of the DoJ from 1965 to 1968, further advanced these ideas by proposing that firms in oligopolistic industries should be broken up if price competition was unlikely to emerge. His advocacy stemmed directly from the belief that concentrated market power inherently leads to anti-competitive outcomes (Hovenkamp 2005).

The Consumer Welfare Approach

By the 1980s and 1990s, a new consensus emerged among antitrust economists, lawyers, and judges: that the primary goal of antitrust law should be to assess whether consumers are actually harmed when firms engage in exclusive deals, merger with rivals, and coordination effort along the supply chain. The transition from the populist approach, centered on broader concerns about market power and its potential political implications, to the consumer welfare approach, prioritizing economic efficiency, marked a significant shift in antitrust law and enforcement. output, consumer prices, and innovation.[126]

[126] This consensus had a monopoly—forgive the pun—on antitrust thinking until around 2016, when a renewed populist critique began to gain traction among academics, activists, and politicians. In the conclusion to this book, I will explore how neo-populists have advocated for a

The consumer welfare approach managed to put several concrete questions front and center. Can a large firm in a concentrated market actually exercise market power by rationing quantity and increasing prices? Can a firm with market power raise barriers to entry, making it less likely for other firms to enter the market and drive down prices? Is the firm using its dominance in one market to monopolize another, using exclusionary practices that suppress competition and keep prices high?

In addressing these questions, courts consistently emphasized that their decisions should be grounded in rigorous economic analysis and empirical evidence. The focus shifted to whether a firm both exercises and abuses market power within a clearly defined market, which must first be delineated in terms of potential substitutes (see DoJ and FTC 2010; Hovenkamp 2005). To arrive at these determinations, courts and enforcement agencies had to make precise assumptions about consumer demand, including its elasticity, and products' average and marginal costs (see Areeda and Turner 1975; Baumol 1982). They had to consider fine-grained data, including prices, profit margins, business behavior, and R&D spending in real markets (see Kwoka and White 2014), while downplaying simplistic concentration measures like the number of firms operating in any given market. As Muris (2023: 59) notes, "By focusing on the welfare of consumers, modern antitrust analysis, including for mergers, uses concentration data as a sometimes-important input, but not an end in itself. Big is neither inherently bad nor inherently good."

Judicial Patterns in the Consumer Welfare Approach

---

return to a broader interpretation of antitrust law that considers issues like choice availability, market concentration, and political power.

During the consumer welfare era, courts tolerated, if not championed, larger firms if they competed vigorously, condoned monopolies that achieved market power through superior business acumen, and adopted a conservative, wait-and-see approach in the face of rapid technological change. In *Ball Memorial Hospital* (1986), writing for the US Court of Appeals, Seventh Circuit, Judge Easterbrook emphasized that injuries to rivals are not inherently antitrust violations but are often the byproducts of vigorous competition. In *Microsoft* (2001), the US Court of Appeals, District of Columbia Circuit ruled that, for them to be judged illegal, monopolistic acts must harm the competitive process and consumers—not merely harm competitors. Going further, in *Verizon Communications* (2004), the Supreme Court underscored that, because it attracts business acumen and fosters innovation and economic growth, monopoly prices are an integral part of the free-market system.

Moreover, during this era, Supreme Court justices across the ideological spectrum showed remarkable consensus in prioritizing consumer welfare and efficiency over the protection of small businesses or similar considerations. In *Verizon Communications* (2004), the Court unanimously emphasized that the mere possession of monopoly power is not unlawful unless accompanied by anticompetitive conduct. Additionally, in *Pacific Bell* (2009), the Court unanimously held that, for price-squeezing claims to be valid under antitrust laws, they must involve specific anticompetitive conduct; merely showing that a firm's aggressive pricing strategy puts its competitors under pressure is necessary, but not sufficient.[127] Similarly, in two

[127] In its lawsuit, LinkLine alleged that Pacific Bell set its wholesale prices appreciably higher than the retail prices it charged its own customers, making it impossible for LinkLine and other competitors to remain competitive in the retail market (see Pacific Bell 2009).

other cases where plaintiffs alleged exclusionary conduct, *Weyerhaeuser* (2007) and *NYNEX* (1998), the Court decided in favor of the defendants sans dissenting opinions.[128]

HOW THE CONSUMER WELFARE STANDARD EMERGED

Why and how did this transition from populism to a more liberal approach to antitrust come about? The conventional view holds that the so-called Chicago School's critiques of the competition policy status quo provided an intellectual foundation that then rapidly infused actual policy (Hovenkamp 2005; Hazlett 1998). Emerging in the mid-20th century and rooted in the University of Chicago's economics department and law school, it advocated for a more rigorous, economics-based approach to antitrust, emphasizing consumer welfare and economic efficiency.

There were several key figures in this movement. Judge Robert Bork contended that though antitrust laws were originally intended to protect consumer welfare, many enforcement cases and legal verdicts perversely harmed consumers: they penalized efficient business practices that could lead to lower prices and improved products (Bork 1978). This is because antitrust actions were often driven by an inherent distrust of big business and not grounded in sound economic principles. Meanwhile, Richard Posner contributed significantly to the empirically

---

[128] *Weyerhaeuser* (2007) dealt with claims of predatory bidding. The Supreme Court ruled that the same stringent standards for proving predatory pricing must apply: a plaintiff must show that the alleged predator bid up input prices to a level that made it impossible for rivals to compete and that the predator would recoup the losses once rivals were driven out. *NYNEX* (1998) involved a claim of illegal tying. The Court held that a change in supplier without an anticompetitive agreement did not constitute a per se violation of the antitrust laws, prescribing that alleged anticompetitive effects associated with this type of business practice must be proven.

grounded economic analysis of law and George Stigler explored regulatory capture and the inefficiencies of government intervention (Hovenkamp 2005; Hazlett 1998).[129]

These scholars and others like them argued that large companies could achieve significant efficiencies, such as economies of scale, which ultimately benefit consumers. They maintained that if a firm achieved a monopoly by offering better products, services, or lower prices, this outcome was not only natural but also beneficial to consumers. And they averred that market forces were generally capable of correcting anti-competitive behavior over time, even in cases in which monopolies reduced consumer welfare, thus obviating the need for heavy-handed government intervention. Specifically, they challenged the structuralist view of "barriers to entry," arguing that factors like high capital requirements or advertising costs were often just reflections of efficiency rather than artificial hurdles to competition (see Stigler 1968). Furthermore, they revolutionized the view of vertical restraints, contending that a firm with a monopoly at one level of the supply chain generally has no economic incentive to "leverage" that power into another level unless doing so creates genuine distribution efficiencies that benefit the end consumer (see Bork 1978).[130]

---

[129] Moreover, although he published much less on these matters, Aaron Director, along with Richard Posner, helped found the law and economics approach and championed the importance of empirical evidence in determining the competitive effects of business practices (see Liu 2022).

[130] While the Chicago School's initial revolution was rooted in neoclassical price theory, its "technocratic" toolkit was later enhanced by the adoption of game theory (pioneered by scholars like Nash 1950 and Milgrom and Roberts 1990). On one hand, it allowed analysts to model how firms might use "sunk costs" or "limit pricing" to deter entrants—addressing concerns about

Scholars supporting the view that the Chicago School's advocacy on behalf of a consumer welfare approach to antitrust *was the actual cause of the shift to that approach* often point to US Supreme Court rulings that cite these authors. To give just one example, when espousing the efficiencies engendered by franchisors that demand non-price restraints on their franchisees, the Court leaned heavily on the work of Bork and Posner in *Continental T.V.* (1977), a case where electronics maker Sylvania restricted its franchisees from selling its televisions at unauthorized locations in a bid to incentivize dealers to provide better service and thus compete more effectively against other brands (see Hovenkamp 2005). It is one thing to espouse these views, however; it is another thing altogether to change the world informed by them.

Challenging the Conventional Explanation for the Shift in Antitrust Policy

While the conventional view outlined above attributes the shift from a populist to a more conservative, consumer welfare-oriented approach to antitrust to compelling Chicago School arguments about the virtues of price theory—more on that below—and economic efficiency (see e.g., Wu 2018; Khan 2017), this explanation oversimplifies a complex, multifaceted transition. Indeed, such an ideational view implies that the Chicago School's critiques of the populist approach to antitrust were so intellectually compelling that they singlehandedly induced a

---

market power that price theory sometimes oversimplified. On the other hand, it provided a more rigorous justification for vertical restraints by showing how "self-enclosed systems" (like those later seen in digital ecosystems) could solve complex coordination and "free-rider" problems that ultimately benefit consumers (see Tirole 1988). In this sense, Game Theory functioned as the "Post-Chicago" software running on the Chicago School's "Consumer Welfare" hardware.

wholesale shift in antitrust policy and judicial interpretation; by the late 20th century they somehow extended beyond academia and infiltrated the judiciary and regulatory agencies.

The historical evidence suggests, however, that the transition towards prioritizing consumer welfare was instead driven by a confluence of factors and voices, many predating the Chicago School's rise to prominence. Indeed, the focus on using antitrust law to benefit consumers has been a central theme across various strands of antitrust analysis since the inception of laws governing competition. Long-standing opposition to the Robinson-Patman Act, for example, demonstrates that critiques of the populist antitrust approach began well before the Chicago School's influence.

Critics such as Fred Rowe, who began his opposition to the Act as a Yale student in 1951 and later published a major treatise on it in 1962, exemplify early challenges to the populist emphasis on protecting small businesses over consumer welfare (Rowe 1962). Economists like MIT's Morris Adelman and Havard's Donald Turner similarly criticized the law's focus on chain stores, arguing in the 1940s that antitrust enforcement should prioritize consumer interests rather than merely shielding competitors. Adelman's comprehensive book on the subject, published a decade later, further solidified this viewpoint (see Adelman 1959).

Prominent legal scholars such as Milton Handler of Columbia and Thomas Kauper of Michigan also argued against the populist tendency to protect competitors rather than competition itself (Handler 1957; Kauper 1981). Additionally, as outlined above, while still aligned with populist prescriptions in certain respects, and whatever its flaws, the structural approach to antitrust championed by Joe Bain and Edward Mason laid the groundwork for a more price-focused and consumer welfare-oriented antitrust approach (see Bain 1956; Mason 1964), well before the Chicago School rose to prominence (see Muris 2023).

Later, Phillip Areeda and Donald Turner were pivotal in shaping modern antitrust thought by rejecting the populist approach while also maintaining a critical distance from the Chicago School (Areeda and Turner 1978). On the one hand, they opposed the populist desire to protect small businesses and individual competitors.[131] Instead, these scholars believed that a firm's size or its market power was not inherently problematic, and they affirmed that antitrust enforcement should prioritize maintaining competitive markets and improving consumer welfare (see Areeda and Hovenkamp 1978). On the other hand, Areeda and Hovenkamp did not fully endorse the Chicago School's strict emphasis on economic efficiency as the sole criterion for antitrust enforcement, acknowledging that while some degree of market power might be essential for fostering innovation and achieving efficiency, when it becomes "excessive" it can culminate in anticompetitive outcomes (see Crane 2007). Consequently, they advocated for a nuanced, case-by-case analysis that balances the goals of promoting economic efficiency with those of preserving competitive market structures (see Kovacic 2007).[132]

---

[131] "Antitrust Law: An Analysis of Antitrust Principles and Their Application" was initially authored by Phillip Areeda and Donald Turner but, following Turner's retirement and Areeda's passing, the multi-volume treatise was maintained and modernized by Herbert Hovenkamp, who remains its primary author today (see Areeda and Hovenkamp 2020).

[132] By the same token, despite the conventional view of the Chicago School movement as a monolithic force for deregulation (see Hovenkamp 2005; Hazlett 1998), its leading figures often disagreed on the levels of market concentration that merited government intervention via merger challenges or monopolization suits. Indeed, cutting against type, scholars like Bork and Posner remained wary of extreme consolidation—such as "four-firm" ratios exceeding 60%—

This gaggle of academic dissenters against the antitrust status quo that emerged during the early post war period was joined by government officials who criticized its incoherence and arbitrariness well before the Chicago School rose to prominence. Notable examples include the reports issued by AGNCSAL, the Attorney General's National Committee to Study the Antitrust Laws (e.g., AGNCSAL 1955), FTC Commissioner Philip Elman's significant dissents in Robinson-Patman cases during the 1960s (see e.g., FTC 1966), and broadsides against the FTC's administrative overreach and "structural rigidities" issued by the American Bar Association in 1956 and 1970 (ABA 1956; ABA 1969; see Muris 2023: 9).[133] Finally, a 1979 special report on antitrust commissioned by the federal government without nary a mention of Chicago School authorities championed prioritizing competition, consumer welfare, and economic efficiency over the protection of small businesses (see NCRALP 1979). Besides, other economic schools of

---

illustrating that the movement's pivot toward consumer welfare was less a total surrender to bigness and more a shift toward a technocratic and empirically guided approach that called for government scrutiny in cases of clear threats to competition (Bork 1978; Posner 2001).

[133] The 1969 ABA report, often referred to as the "Kirkpatrick Report" (after its chair, Miles Kirkpatrick), remains one of the most significant institutional critiques in antitrust history. Commissioned by President Nixon at the urging of the ABA, the report was a scathing indictment of the FTC's "structuralist" obsession with trivial cases, such as small-scale Robinson-Patman price discrimination disputes. It famously warned that if the FTC did not replace its populist-driven enforcement with a more rigorous, economics-based approach—prioritizing "matters of significance" over "the protection of small competitors"—it should be abolished altogether (see ABA 1969; Muris 2023).

thought besides the Chicago School also significantly influenced antitrust between the 1970s and 2000s.

Insights by economists like Ronald Coase (1937) and Oliver Williamson (Williamson 1975; 1985), which revealed the inner workings of firms and explained their business strategies, reshaped antitrust decision-makers' views on practices like vertical integration and exclusive dealing. For example, this so-called industrial organization approach showed that, rather than merely restricting market access or harming competitors, vertical non-price restraints such as exclusive territories and location clauses could align incentives between franchisors and franchisees and thus prevent freeriding, ensure better customer service, preserve brand reputation, and promote interbrand competition (see Williamson 1985; Hovenkamp 2005). Additionally, Coase and Williamson both argued that vertical integration is almost always efficiency-enhancing rather than anticompetitive, challenging the populist notion that vertical mergers were inherently exclusionary (see Brown Shoe 1962). Accordingly, in the landmark Continental T.V. (1977) case, the Supreme Court ruled in favor of GTE Sylvania's territorial restrictions on its dealers. Additionally, enforcement agencies adopted a more receptive stance toward vertical integration in antitrust enforcement (see DoJ 1984). (see DoJ 1984).

Additionally, ideas about dynamic efficiency from Schumpeterian economics played a crucial role in reorienting antitrust law and enforcement towards promoting innovation rather than merely focusing on output and prices in the short run. Schumpeter (1942) argued that large firms with significant market power are often best positioned to innovate, driving economic growth through a process of "creative destruction." The major effect this had on antitrust was to open the door to a greater tolerance to firms achieving bigger sizes and experimenting with

mergers and business strategies that might endow them with greater market power but improve consumer welfare in the process (see Teece 1984; Katz and Shelanski 2007).

Notably, this shift in perspective started well before the Chicago School's advocacy for the consumer welfare standard. During the late 1940s, scholars like Donald Turner critiqued the antitrust prosecution of A&P for its vertical integration strategy. Turner argued that "the lure of temporary monopoly profits is an important impetus to the introduction of new products and new techniques, which rudely upset the peaceful, profitable existence of long-entrenched business firms. This constant change to the new, the more efficient, is the very heart of the process of effective competition" (see Turner 1949: 1969-971). Ward Bowman also contributed to this evolving understanding (see Bowman 1957). He challenged the traditional view that tying arrangements were inherently anticompetitive, instead arguing that such practices could enhance efficiency by enabling firms to better compete and innovate, thus benefiting consumers in the long run.

In short, the idea that the Chicago School alone gave rise to the consumer welfare approach to antitrust, ultimately fostering a more permissive environment for Schumpeterian creative destruction and the rise of digital platforms, is not accurate.

Organic Ascendance of the Rule of Reason Approach

A better explanation for the rise of the consumer welfare approach to antitrust is that justices and policymakers striving to interpret the language and logical implications of the Sherman Act, Clayton Act, and Robinson-Patman Act pragmatically, and gradually, honed a Rule of Reason approach. Trying to *arrive at the facts* contested in lawsuits forced them to *assess the actual competitive effects of business practices* on competition and consumer welfare in *real life market scenarios*. While recognizing that each case required individualized judgment, they also

encountered recurring patterns. And it is only natural that various economic approaches, including the version of price theory advanced by the Chicago School, but all the others I outlined above too, provided antitrust decision-makers with valuable frameworks to make sense of these facts and patterns. Therefore, I submit that the real cause of the consumer welfare approach to antitrust is organic evolution of the Rule of Reason standard for judging firm behavior as a technical exercise in its own right that preceded the Chicago School's advocacy on its behalf.

By the late 20th century, the Rule of Reason had transformed from a vague judicial "sniff test" into an exacting monopoly detecting machine. This shift from intuition to technocracy was not merely an ideological preference for laissez faire economics, but a structural necessity. Once courts demanded proof of "harm to competition" rather than just "harm to competitors," policymakers were compelled to operationalize antitrust law through microeconomic fundamentals.

The evolution of antitrust was not a straight line from "populist fear and angst" to "economics-informed technocracy," however. Instead, it was a complex, non-linear journey away from per se prohibitions—where certain conduct like price-fixing or group boycotts was "plainly anticompetitive" and prohibited without further inquiry (see *Socony-Vacuum* 1940)—toward a dominant Rule of Reason.

Irrebuttable Illegality

The per se rule was centered on the "nature of the restraint" itself rather than its specific market consequences; once a practice was categorized as falling within a prohibited class, it was deemed irrebuttably illegal without regard for any purported business justifications (see Northern Pacific Railway 1958; Hovenkamp 2005). Historically, this categorical approach saw a resurgence in the

mid-20th century as the Supreme Court sought judicial "bright lines" to manage an increasingly complex industrial economy. This era established rigid prohibitions against price-fixing (Trenton Potteries 1927), market division (Topco Associates 1972), and vertical price maintenance (Dr. Miles Medical 1911). By the 1950s and 60s, the Court had extended per se illegality to group boycotts (Klor's 1959) and tying arrangements (Northern Pacific Railway 1958), effectively declaring a wide swath of corporate strategies illegal regardless of their actual impact on output, prices, or innovation.

While the per se rule offered judicial efficiency, it often resembled a blunt instrument; as business strategies became more "variegated," the legal system encountered a recurring friction: rigid rules were frequently punishing conduct that actually improved market efficiency and seemed to improve consumers' lives. This friction led to a series of fitful "advances and retreats and subsequent advances": a seesawing between the Rule of Reason agnosticism and per se restrictions.

Rule of Reason's Evolution

The conceptual roots of the Rule of Reason are English common law's notion of reasonableness—most notably Mitchel (1711). In that case, Lord Macclesfield established a distinction between "general" restraints of trade (which were void) and "specific" or "ancillary" restraints—such as a baker's agreement not to compete within a local parish for a set time—which were enforceable if they were reasonable and based on valid consideration.

This principle of "ancillary" restraints was later adopted by American courts, as seen in Oregon Steam Navigation (1874) and Judge Taft's influential opinion in Addyston Pipe & Steel (1898), which distinguished "naked" price-fixing from restrictions that were merely incidental to a legitimate business transaction. Chief Justice Edward White's landmark opinion in Standard

Oil (1911), which formally established that the Sherman Act did not ban all restraints of trade, but only those that were "unreasonably" restrictive. This opened the door for the "Chicago Board of Trade" (1918) decision, where Justice Brandeis—ironically, given his later populist stance—argued that the legality of an agreement cannot be determined by a simple label, but must instead turn on the history, nature, and effect of the restraint.

And while the Supreme Court briefly retreated into a rigid per se rule against vertical restrictions in Schwinn (1967) and also did so in 1966 in the case of Von's Grocery, where the Court focused on mere market concentration to block a merger that actually promised significant consumer efficiencies, it decisively pivoted back to a nuanced, economically informed framework in Continental T.V. (1977). This was not an ideological surrender to the Chicago School; rather, it was a pragmatic realization that the "machinery of economics" was the only tool capable of distinguishing a "naked" restraint from "ancillary" efficiency (see also Addyston Pipe 1898).

But What is the Rule of Reason, Exactly?

The Rule of Reason is a judicial doctrine used to assess whether a firm's conduct violates antitrust laws based on its competitive effects rather than the specific nature of the conduct itself. By weighing pro-competitive effects against anti-competitive ones, this approach requires courts and enforcement agencies to conduct a thorough analysis of the behavior of specific firms within a particular market context on a case-by-case basis. It generally follows a burden-shifting framework: the plaintiff, whether a government entity or a private party, must first demonstrate that the defendant's conduct has a significant anti-competitive effect within the relevant market. Once this burden is met, the defendant is given the opportunity to present pro-competitive justifications for its conduct, such as efficiency gains or innovations that benefit consumers. The

court or enforcement agency then balances these anti-competitive effects against the pro-competitive justifications, ultimately determining whether the conduct is lawful based on whether the benefits outweigh the harm (see, e.g., Standard Oil 1911; Chicago Board of Trade 1918; Reiter 1979; California Dental Association 1999: 770). In both monopolization and merger cases, the Rule of Reason analysis always begins by defining the relevant product and geographic market in which a firm is alleged to hold market power.

In monopolization cases, the analysis then moves to identifying the firm's market power within that relevant market: specifically, whether the firm can unilaterally restrict output and raise prices above the competitive level without losing so many sales that the price increase becomes unprofitable (see DoJ 1982). To determine a firm's market power, enforcement agencies and judges may consider factors such as market shares, profit margins, barriers to entry, and the overall dynamics of the market (e.g., DuPont 1956).[134]

In merger cases, after establishing market boundaries, enforcement agencies and courts assess the market's current level of concentration and speculate about how the merger might alter this structure. They speculate about whether the new, post-merger firm might go on to possess significant market power, potentially enabling it to reduce output, raise prices, decrease quality,

[134] While a market share of 70% is generally required to make a prima facie case of market power, shares between 40-70% may also be sufficient when considering additional factors. Moreover, a firm's ability to price discriminate may constitute an indirect test of market power (see Grinnell 1966; Areeda and Hovenkamp 2023). Additionally, the Rule of Reason is a flexible standard that may allow for an abbreviated "quick look" analysis when the anticompetitive nature of a restraint is apparent even without a full market study (see NCAA 1984).

or slow down innovation. As with monopolization cases, the merging firms can present evidence of the merger's pro-competitive effects. These may include efficiencies such as economies of scale, which could benefit consumers through lower prices, enhanced products or services, or accelerated innovation. If the merger's benefits to consumers and competition outweigh the potential harm, the merger may be approved. Conversely, if the harms are deemed too significant, the merger may be blocked or approved only after the federal government imposes consent decrees—legal agreements that require the companies to accept "fixes" such as structural divestitures (selling off parts of the business to a rival) or behavioral conduct requirements (rules dictating how they must treat competitors)—which aim to mitigate anti-competitive effects such as unilateral price increases or coordinated effects, where the few remaining firms in a market stop competing and quietly raise prices together (DoJ 1982).

Importantly, for our purposes, in multisided markets—platforms—enforcement agencies and courts define multiple relevant markets to reflect each side of the platform. This approach respects the microeconomic logic of platforms, which relies on the dynamics of indirect network effects and differential pricing schemes across the different sides of the market (see Evans and Schmalensee 2016; Wright 2004; and Yun 2019).

When enforcement agencies and courts identify the multiple relevant markets that represent the different sides of a platform, they are better able to identify the full scope of costs and benefits of firm behavior and engage in more accurate balancing tests. For example, if consumers use a service for free on one side of the market while advertisers pay to access these consumers on the other side, analyzing only the consumer side might suggest the platform has no market power. However, if the platform collects substantial fees from advertisers, such a judgement may prematurely rule out competitive concerns. Conversely, by providing consumers

with free access to the platform, the firm may enhance user engagement and juice data generation in ways that improve service quality, attracting more advertisers that chase more users and that ultimately reduces the fees charged to advertisers (see Evans and Schmalensee 2016). Ultimately, if the benefits provided to consumers are substantial, antitrust decision-makers may condone "high" prices on the advertiser side of the market (see American Express 2018).

The Rule of Reason Contained the Seeds of Consumer Welfare Approach

How did the Rule of Reason come to dominate antitrust analysis? Practical necessities led jurists and policymakers to gradually incorporate economic principles like price theory into their decisions. This evolution was neither premeditated nor ideologically driven; rather, it emerged organically as courts and enforcement agencies grappled with real cases involving alleged consumer harm. Antitrust decision-makers didn't require advocates to remind them of the importance of focusing on market prices rather than fixating on firm size or market concentration as such: they were already inclined to adopt a more economically informed approach well before the Chicago School emerged on the scene.

To be sure, the Chicago School honed price theory and other microeconomic tools and advocated for their full embrace by antitrust decision-makers under the consumer welfare standard (see Liu 2022)—however, it was the judiciary that initially grappled with these economic concepts. Faced with tangible cases that pitted adversaries against each other, judges had to work through the internal logic of real-world allegations about business conduct and its purported effects. To make sense of conflicting claims made by plaintiffs and defendants about the same evidence, courts had nowhere else to turn but the machinery of economics.

Antitrust decision-makers soon recognized that to assess whether a firm possesses market power—and can duly exploit it to substantially reduce competition—they needed to first define

the relevant market, as the source of market power ultimately lies in a firm's ability to manipulate output, prices, and investment. Crucially, "[b]ecause courts may trade off pro- and anticompetitive effects within the relevant market, defining the relevant market is tantamount to establishing the space of allowable trade-offs" (Ward 2017: 2078).

This realization naturally led enforcement agencies and courts to adopt economic tools capable of identifying and measuring a relevant market and, by extension, a firm's power within this market. These tools, in turn, laid the groundwork for what would later be known as the consumer welfare standard. Defining a relevant market requires identifying and calculating the substitutability of products or services, as only products perceived as interchangeable by consumers belong to the same market. What products serve as substitutes for the allegedly monopolized product? Would a merger between producers of these substitutes create a new monopoly?

Price theory, with its focus on supply and demand curves and equilibrium prices, was uniquely equipped to provide antitrust decision-makers with the answers to these questions. By prioritizing output levels as the objective metric of competition, it offered a way to distinguish between a firm that was simply "too big" and one that was actively harming the market by restricting supply to inflate prices. Price theory was ultimately a very practical framework that helped judges and enforcement agencies to quickly and decisively define markets and assess the competitive effects of firm behavior within them.

First, price theory established that a firm supplying a product with few or no substitutes wields significant market power. This led antitrust decision-makers to devise ways to measure substitutability through factors such as cross-price elasticity, diversion ratios, and consumer preferences. Cross-price elasticity gauges the responsiveness of demand for one good when the

price of another changes.[135] Diversion ratios capture the proportion of consumers who switch from one product to another following a price increase.[136] While rooted in academic theory, these metrics were eventually codified into the federal government's Horizontal Merger Guidelines, transforming the subjective "judgment calls" of early antitrust into a standardized, technocratic process (see Shapiro 1996; DoJ and FTC 2010).

Second, price theory offered practical tools to assess competitive dynamics and the consumer impacts of business practices, including concepts like consumer surplus and the Lerner Index for pricing power and profit margins. Consumer surplus—the difference between what a consumer is willing to pay for a product and the actual price paid—represents the benefit derived from market participation. As I explained in Chapter 4, the Lerner Index measures a firm's

---

[135] The formula for cross-price elasticity is $\eta_{xy} = \left(\frac{\partial Q_x}{\partial P_y}\right) \times \left(\frac{P_y}{Q_x}\right)$ where Q_x is the quantity demanded of good X, P_y is the price of good Y, and $\frac{\partial Q_x}{\partial P_y}$ is the partial derivative of Q_x with respect to P_y; or, in more intuitive terms, the percentage change in quantity demanded of good X divided by the percentage change in price of good Y.

[136] Diversion ratios work in tandem with the Lerner Index. Recall that the Lerner Index ($L = \frac{P-MC}{P}$) captures a firm's pricing power as the margin of price over marginal cost. In merger analysis, this margin determines the profitability of "recaptured" sales: when a merged entity raises product $A's$ price and some customers switch to product B (now under common ownership), the merged firm earns margin m on each diverted unit. The Upward Pricing Pressure framework developed by Farrell and Shapiro (2010) formalizes this relationship, with the product of the diversion ratio and the margin serving as the key indicator of merger harm.

pricing power by comparing the difference between a product's price and its marginal cost, with a higher index indicating greater market power.[137]

Third, and logically implied by the points made above, price theory allowed analysts to differentiate between true market power and simple competitive advantage. By capturing the elasticity of the supply curve, it reveals whether a firm—though a "price taker"—earns higher profits simply because it exhibits lower marginal costs than its rivals. This helps distinguish Ricardian Rents (profits earned through superior efficiency) from monopoly rents earned by inducing scarcity (see Chapter 4)—instrumental for knowing whether "extraordinary" profits are a "natural" result of superior processes or, instead, a signal that barriers to entry have short-circuited the competitive process.[138]

Fourth, price theory integrated insights regarding product differentiation to further refine this distinction. When consumers willingly pay a premium for superior quality, branding, or innovative features, the resulting price difference often reflects value creation rather than a lack of substitutes (see Porter 1985; Galetovic 2021 in the case of semiconductors). Courts reinforced this technical distinction by ruling that the introduction of a technologically superior product—

---

[137] As I stated in Chapter 4, the Lerner Index formula is $L = \frac{P - MC}{P} = \frac{1}{|\varepsilon|}$, where P is the price, MC is the marginal cost, and ε is the elasticity of demand. In the legal arena, this formula provided the foundation for the "Areeda-Turner test," a highly influential rule adopted by courts to distinguish legitimate price competition from illegal predatory pricing based on whether a firm's price fell below its average variable cost (see Areeda and Turner 1975).

[138] This distinction informed several landmark cases where courts had to decide if high profits were a "reward" for excellence or a "penalty" on consumers (e.g., *Eastman Kodak* 1992).

even if it makes a rival's offerings incompatible—is a pro-competitive benefit rather than an illegal tie-in (see Foremost Pro Color 1983). However, analysts also used price theory to identify when such differentiation was being leveraged as a barrier to entry—using unique product attributes to deter new competitors and maintain dominance (see Tirole 1988).

Fifth, and finally, antitrust analysts turned to price theory to understand the dynamics governing vertical relationships—the links between a firm and its suppliers or distributors. By analyzing factors like bargaining power and supply elasticity, judges could identify how much economic value a firm added to a supply chain (Syufy 1990). Firms that control critical technological components or possess significant brand strength add more value and thus earn higher margins (see Porter 1980). Courts learned that a firm's leverage over upstream suppliers or downstream retailers often stemmed from its strategic importance in the product's supply chain, which implied that vertical arrangements that appeared exclusionary could have legitimate efficiency justifications and should not be reflexively condemned (see Williamson 1975; Klein et al. 1978).

These insights—honed through decades of case-by-case application—provided the technical scaffolding for a consumer welfare standard long before it was given a formal name. To explain how these technical components functioned in practice, I review some landmark court decisions and enforcement agency tools that grappled with fundamental questions about market boundaries and market power.

Grappling with Price Theory

Long before the Chicago School to antitrust was in diapers, in Alcoa (1945) the Second Circuit positioned the "market definition" problem at the center of antitrust analysis. Judge Learned Hand recognized that identifying monopoly power was impossible without first delineating the

product's geographic and functional scope—a process that required a technical evaluation of the demand curve and the cross-elasticity of demand. In this case, the court was forced to decide whether the market was limited to primary aluminum or extended to "secondary" (recycled) producers. This wasn't merely a boundary dispute; it was a first-of-its-kind attempt to use price theory to assess whether Alcoa could unilaterally restrict supply to inflate prices.

Judge Hand's decision famously moved the inquiry away from populist questions of "fairness" and toward the mechanical reality of supply control. He concluded that the ultimate metric of market power was not a firm's size, but its objective ability to dictate terms: "The test of success is not whether the applicant has obtained a price which is economically fair, but whether it is able to dictate the price by reason of its control of the supply" (Alcoa 1945: 427). As Hovenkamp (2005) and Posner (2001) have noted, this distinction was a watershed moment; it effectively grounded the judicial inquiry in the objective logic of price theory—specifically the elasticity of supply—rather than the subjective whims of moral sentiment. In this sense, Alcoa functioned as the first major prototype for a "monopoly detection machine" calibrated to detect economic effects rather than social "bigness."

Judge Hand's technical approach influenced subsequent debates over how to prevent judges from overreaching when defining markets. In his dissenting opinion in Continental Can (1964), Justice Harlan—echoing Hand's requirement for empirical rigor—criticized the majority for adopting an arbitrary market definition that conflated functionally distinct types of packaging (glass and metal) into a single market. Harlan argued that this approach ignored real competitive dynamics, famously accusing the majority of "inventing" a line of commerce "the existence of which no one, not even the Government, has imagined" (Continental Can 1964: 476-77).

This dissent highlighted the technical pitfalls of overly broad market definitions, which—if not grounded in the cross-elasticity of demand—inevitably lead to incorrect conclusions about a firm's actual market power and its effect on competition (see Areeda and Turner 1978). This "imaginary" boundary-setting was the exact error that would soon manifest in the infamous DuPont decision.

However, this is where we reach an inflection point: While the Supreme Court correctly identified "reasonable interchangeability" in DuPont (1956) as the gatekeeper for market power, it mistakenly ruled that the DuPont company did not possess a monopoly because it defined the market too broadly to include all flexible packaging materials. This decision infamously introduced the Cellophane Fallacy—where the Court evaluated potential substitutes at *current monopoly prices* rather than competitive ones (see Stocking and Mueller 1955). In this case, the Court observed that consumers could switch from Cellophane to products like wax paper and aluminum foil, inferring that Dupont lacked a monopoly, but it failed to recognize that these were "inferior" alternatives that only appeared viable because of DuPont's artificially inflated prices (see Areeda and Hovenkamp 2003).[139]

As a result, the court's broad market definition underestimated DuPont's true market power. This taught the "monopoly detection machine" a vital lesson: identifying market power first requires evaluating what substitutes would look like at competitive price levels—where marginal revenues equal marginal cost—to approximate the true cross-elasticity of demand (see

[139] At this price point, a monopolist cannot increase prices further in a way that is profitable; doing so causes consumers to exit the market, shrinking the monopoly rents earned on remaining sales.

Elzinga and Hogarty 1973; Kaplow 2010).[140] In other words, because the real world betrayed a survivorship process that distorted competitive benchmarks, in that consumers' behavior adjusts to the revealed market structure's equilibrium quantity and prices—the relevant market had to be defined counterfactually. The inquiry shifted to what the market *would have looked like* if purchasers could not divert their consumption into potentially inferior substitutes. Courts soon internalized this lesson, moving toward a more rigorous technical assessment of whether observed consumer switching patterns in fact signal the absence of a firm's market power or instead *demonstrate its successful exercise*.

Enforcement Agencies Operationalized These Insights

The subsequent development of the modern antitrust toolkit was not a mere academic indulgence in formal modeling; it was a practical response to a systemic policy vacuum. As the judiciary shifted the Rule of Reason's focus from "fairness" to "efficiency," antitrust practitioners found themselves in a high-stakes dilemma: they were legally required to prove competitive harm but lacked the empirical instruments to measure it. This meant that practically minded economists responded, defining a market scope's by identifying ex ante demand curves and measuring market power not by a firm's size or intentions, but by its pricing power as dictated by actual demand cross-elasticities.

Foremost among the tools antitrust practitioners relied upon to properly define a market is the Hypothetical Monopolist Test (HMT). Rather than recommit the "Cellophane Fallacy" by starting from a potentially distorted status quo where firms may be earning monopoly rents even

[140] Kaysen (1956) demonstrates that DuPont's prices indeed satisfied the criteria of monopoly pricing according to the Lerner Index.

though consumers appear to consume substitute goods, this methodology establishes a true competitive price benchmark (see Werden 1992; Werden 2003; Baker 2007). As Gregory Werden (1992)—one of the primary architects of the DoJ's 1982 Merger Guidelines, where the HMT first appeared (see DoJ 1982)—observed, this tool was specifically engineered to solve the "lack of a clear, workable, and economically sensible" method for market definition that had plagued previous decades of litigation, thereby providing a "principled basis" for determining which products actually constrain one another's pricing power (Werden 1992: 139).[141]

The HMT resolves the "broadening" puzzle associated with the "Cellophane Fallacy" by working from the bottom up. It begins with the narrowest possible "candidate market", in this case, a single product, and asks: If a hypothetical monopolist raised the product's price by 5%—a small but Significant and Non-transitory Increase in Price (SSNIP)—would it be profitable to do so? If the answer is no, because consumers would switch to a specific outside alternative, that alternative is added to the "candidate market," and the test repeats. If the answer is yes, the expansion stops immediately and the market is defined as containing only those products (see DoJ 1982).

Note that this exercise would naturally exclude the inferior products to Cellophane, such as wax paper, which were mistakenly included in the DuPont market boundary setting (see DuPont 1956). At a competitive price baseline, the qualitative gap between the primary product and its inferior alternatives is too wide for a mere 5% price increase to trigger significant consumer switching, meaning the hypothetical monopolist could profitably raise prices without "leaking" demand to the inferior substitute. However, at some higher thresholds—say, a 10% or

[141] The FTC also adopted these guidelines (FTC 1982).

20% price increase—consumers would indeed switch to inferior products, and it would artificially look like the market is broader than it actually is. By identifying the hypothetical price increase by the hypothetical monopolist in the "right neighborhood," the HMT is can effectively capture the cross-elasticity of demand between true substitutes and therefore identify the smallest group of products that truly constrain one another (see Kaplow 2010; Baker 2007).

Consider a hypothetical soft drink merger. If a 5% price hike on Colas causes consumers to switch to flavored seltzer water, the aggregate elasticity of demand for soft drinks is relatively high, and seltzer must be included within the market boundary since it acts as a genuine competitive check: in the eyes of consumers, it is a true substitute to Cola, providing them with a comparable product of equal quality. Conversely, if antitrust authorities were to unduly exclude seltzer, they would commit a “Narrowing Error,” artificially tightening the market and misidentifying a competitive price-taker, the Cola company, as a monopolist.

In other words, the HMT avoids the Cellophane Fallacy by ensuring that the market’s boundaries only expand based on consumer behavior at competitive prices. While the DuPont court mistakenly included inferior products that only appeared as substitutes because prices were already at monopoly levels the HMT "cleanses" the analysis by starting the iteration at a competitive baseline. The test ensures that the resulting market definition is not too broad (a result of a monopolist diverting consumption to inferior products by jacking up the price of a high-quality good) but is instead calibrated by the true demand cross-elasticity between products, be they Cola and flavored seltzer water or whatever else.

The implication of the HMT is profound. By iteratively expanding the candidate market only until reaching the point of profitability, it established a logically grounded, empirically verifiable instrument to delineate market boundaries and thus bound the demand curve. This

allows antitrust authorities to evaluate whether a firm actually exerts market power by rationing quantity and increasing price—or if the perceived "power" is merely an illusion. By adopting this tool, antitrust crossed the Rubicon from a populist crusade into a technocratic exercise: the term "monopoly" was no longer a moral judgment, but a quantitative threshold defined by whether a firm could sustain margins above marginal cost in a market insulated from competitive constraints.

While the HMT provided the foundation, Critical Loss Analysis (CLA) offered a more versatile "stress test" for markets characterized by high product differentiation. Unlike the HMT, which focuses on price increases, CLA allows analysts to assess the so-called critical loss threshold: the maximum percentage loss of sales a hypothetical monopolist could endure while still finding a price increase profitable. Specifically, this threshold is the “break-even point” where revenue lost from decreased quantity equals the additional revenue gained from the higher price (often expressed as $CL = \frac{t}{t+m}$, where t is the price increase and m is the gross margin). If the actual loss in sales—estimated via econometric models or diversion ratios—is less than this "critical loss," the price increase is profitable (see Harris and Simons 1989; Katz and Shapiro 2003).

By comparing this threshold to anticipated sales losses—measured through the cross-elasticity of demand—CLA can identify true market boundaries in complex industries where brand loyalty or "ecosystem integration" might otherwise mask a firm's pricing power. In effect, CLA operationalizes the HMT's central question: if the predicted loss falls short of the critical threshold, the candidate market is properly defined; if it exceeds it, the market must be expanded to include additional substitutes.

Consider the same Cola example used to illustrate the HMT. Suppose a hypothetical Cola monopolist has a gross margin of 40% and is contemplating a 5% price increase. Applying the CLA formula, the critical loss threshold is CL = $\frac{0.05}{0.05 + 0.40} = 11.1\%$. This means the monopolist can lose up to 11.1% of its sales to substitutes like flavored seltzer water and still find the price increase profitable. The next step is to estimate, via econometric analysis or diversion ratios, how many consumers would actually switch. If the predicted loss is 8%—below the 11.1% threshold—the price increase is profitable and Cola alone constitutes a relevant market. If the predicted loss is 15%—above the threshold—the price increase is unprofitable and the market must expand to include seltzer, just as the HMT would require. Both tests thus answer the same fundamental question: can a hypothetical monopolist profitably impose a SSNIP? The difference is methodological: The HMT iterates by adding products until profitability is reached; CLA provides a precise numerical threshold, derived from the firm's own margin, against which predicted consumer behavior can be directly compared.

Once the "monopoly detection machine" had established market boundaries, antitrust authorities moved on to the next logical step: a standardized dashboard to measure the actual depth of a firm's power within a clearly defined space. Drawing on price theory, industrial organization analysis, and game theory, the agencies moved beyond a firm’s mere size to quantify its ability to restrict output and inflate prices. While the primary "dial" on this dashboard is the Lerner Index, antitrust authorities also retrofitted the HMT and Critical Loss

Analysis (CLA) to help the identify market power, as well as calling on new measures centered on “diversion ratios” (Shapiro 1996).[142]

As explained in Chapter 4 and repeated in this chapter, the Lerner Index measures the gap between price and marginal cost and thus directly captures the economic definition of market power. By the late 20th century, this metric became a judicial staple, used in IBM (1969) and Eastman Kodak (1992) to assess pricing power in computers and photocopier parts, respectively, and in the landmark Microsoft case to evaluate market power in the software industry (see Microsoft 2001).

To put the latter-day, post Lerner Index tools in perspective, consider the landmark Staples (1997) case, where the D.C. District Court famously prevented the Staples office supply company from merging with its rival, Office Depot. First, the court relied heavily on the HMT and detailed pricing data to define the relevant market, not as "office supplies" generally, but specifically as "office supply superstores." The economic analysis that took center stage during the suit demonstrated that Staples, Office Depot, and Office Max exerted a unique price-constraining pressure on one another that general retailers like Walmart did not (see ibid). Second, CLA simulated post-merger scenarios, allowing the court to predict with mathematical

[142] Meanwhile, a relic from the “structural era” of antitrust, the Herfindahl-Hirschman Index (HHI), was adopted by antitrust practitioners to measure market concentration by summing the squares of market shares. While in NCAA (1984) the Supreme Court inferred market power directly from high market shares under the assumption that entry was restricted (see Gavil et al. 2014: 185-87), in Dentsply International (2005) the court maintained that a high HHI value was empirically correlated with a monopolistic ability to exclude rivals.

precision whether a merged entity would have the ability to raise prices without triggering a fatal loss in sales (see Moresi and Shapiro 2012). Consequently, the court blocked the merger, concluding that the removal of this specific competitive constraint (Office Depot) would likely allow a Staples-Office Depot composite to appreciably raise prices on office supplies (see Sullivan and Grimes 2006; Kwoka and White 2014).

Moreover, beyond predicting merger effects, these tools were also adapted to measure existing market power in unilateral conduct cases (see Shapiro 1996; Werden 1998). In Qualcomm (2019), for example, the court utilized diversion analysis to measure the absolute depth of Qualcomm's leverage in the market for "smartphone modem chips." By quantifying switching patterns, the court found that Qualcomm faced nearly zero "actual loss" from price increases because its rivals were technically incapable of acting as competitive constraints.[143]

Grounding Monopolization in Economics

Over time, antitrust focus shifted: from merely identifying firms with substantial market power, to understanding how that power was obtained, maintained, or abused through

[143] In merger analysis, diversion ratios measure how much profit the merged entity would recapture when customers switch between its newly combined products. In unilateral conduct cases like Qualcomm, the logic differs: there is no merger and no recapture. Instead, diversion analysis measures whether effective substitutes exist at all. If rivals capture a significant share of diverted sales, the defendant faces a competitive constraint; if diversion to rivals approaches zero, the defendant can raise prices without losing meaningful volume. The Lerner Index then quantifies the extent to which the firm is already exercising this unconstrained power through elevated margins.

exclusionary conduct (see Elhauge 2003; Hovenkamp, 2005). Courts gradually surmised that the Sherman Act's use of the term "monopolize" seemed to imply not just the possession of monopoly power, but the active and willful acquisition or maintenance of that power through exclusionary or predatory conduct (see Elhauge 2003). By extension, this "textualist turn" helped switch emphasis from punishing firms for their success to protecting the competitive process itself (Bork 1978; Landes and Posner 1981), creating a legal shield for the "superstar" firms I examined in Chapter 4: even if they crank production way up to grab more market share, or achieve demand side scale economies, or price well above their marginal cost, they should not face antitrust liability if this was all a result of superior business acumen or a better product.

Practically speaking, antitrust practitioners began to differentiate between types of monopolistic behavior, viewing some firm behaviors that were previously castigated as benign and others as patently harmful to competition or consumer welfare (Areeda and Turner 1975; Posner 1976; Kaplow 1987; Salop 2010). A series of high-profile legal cases codified the idea that high degrees of market power were necessary, but not sufficient, to effectively prosecute and prove a monopolization case (Landes and Posner 1981).

Most seminally, the Supreme Court's decision in *Grinnell* (1966) established that monopoly power alone is not illegal; it must be accompanied by exclusionary conduct that harms competition. The Court clarified that "[t]he offense of monopoly under Section 2 of the Sherman Act has two elements: (1) the possession of monopoly power in the relevant market and (2) the willful acquisition or maintenance of that power as distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident" (ibid: 570-571). Moreover, it emphasized that "[i]t is not unlawful for a company to dominate a market by virtue of a superior product, business acumen, or historic accident; but it is unlawful to maintain that

position through anti-competitive conduct" (ibid: 571). Other landmark cases that followed Grinnell underscored that antitrust laws' aim was to protect the competitive process, not individual competitors, and that enforcement should focus on conduct that harms competition, not the mere existence of market dominance.

In *Spectrum Sports* (1993: 458), the Supreme Court stated that "[t]he purpose of the [Sherman] Act is not to protect businesses from the working of the market; it is to protect the public from the failure of the market" (see also Werden 2015). The Court further clarified that "the possession of market power alone does not violate Section 2; rather, the statute targets the willful acquisition or maintenance of that power as distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident" (ibid: 456).

Similarly, in Rambus (2008: 464), the Supreme Court noted that if a practice "raises the price secured by a seller" or otherwise harms customers but does so without harming competition, it is beyond the antitrust laws' reach (see also Werden 2009); and, in NYNEX (1998: 136) it held that there was no Sherman Act violation where "consumer injury naturally flowed not so much from a less competitive market . . . as from the exercise of market power that is lawfully in the hands of a monopolist . . . combined with a deception worked upon the regulatory agency that prevented the agency from controlling [the monopolist's] exercise of its monopoly power" (see also Elhauge 2003).

Finally, a slew of high-technology centric cases, which I explore below, went further still: they embraced a full-throated, Schumpeterian defense of monopoly. They averred that the lure of rents is what induces the type of risk-taking that drives innovation (see Schumpeter 1942; Chapter 4 of this book).

Ascendance of the Innovation Concern

By the 2010s, many antitrust practitioners had come to embrace various aspects of Schumpeterian economics (see Sidak and Teece 2009; Katz and Shelanski 2005; Evans and Schmalensee 2001). Courts and antitrust authorities acknowledged that market power might be a natural byproduct of innovation and that innovation should be rewarded, not penalized. They also argued that, because market power is often transient in the face of perpetual technological disruption, a lighter regulatory touch is sometimes warranted (see Sidak and Teece 2009; Shelanski and Sidak 2001).

As the Supreme Court incisively articulated in *Verizon* (2004: 398), "[t]he opportunity to charge monopoly prices—at least for a short period—is what attracts 'business acumen' in the first place; it induces risk-taking that produces innovation and economic growth." And "[t]he mere possession of monopoly power, and the concomitant charging of monopoly prices, is not only not unlawful; it is an important element of the free-market system" (ibid: 407).[144] Similarly, quoting *Foremost Pro Color* (1983: 545–46), the Ninth Circuit declared in *Allied Orthopedic Appliances* (2010: 991) that the ability to earn monopoly rents incentivizes economic dynamism.

[144] This case also highlighted the limitations of using antitrust laws to mandate access to a monopolist's infrastructure or technologies, effectively narrowing the "essential facilities" doctrine that was once a staple of populist enforcement (see Areeda 1989; Hovenkamp 2005: 248). The Supreme Court ruled that Verizon's "refusal to deal" with its competitors did not constitute an antitrust violation; rather, it was a legitimate business decision that promoted competition and efficiency and, furthermore, compelling firms to share their infrastructure with competitors could undermine incentives to invest in important improvements and innovation.

This type of reasoning aligned with a more cautious approach to antitrust enforcement in general, particularly in rapidly evolving high-technology markets, where the risk of false positives—erroneous condemnations of procompetitive conduct by antitrust decision-makers—could easily stifle innovation (see Easterbrook 1984: 5). Circumspection of this sort marked a significant departure from the interventionist stance taken towards the big tech firms of the 1960s and 1970s (see Baker 1997: 177-196). For example, explicitly calling out Kodak as the most dominant player in the photography industry in Berkey Photo (1979), the Second Circuit suggested that the company had a duty to widely disclose information about its innovations to give its competitors "a fair opportunity" to compete on a more even playing field. Similarly, in the mid-1970s, to reduce its market dominance, the FTC saddled Xerox with a consent decree that required the copier machine maker to license its xerography technology to competitors (see FTC 1975).

The rising caution towards antitrust intervention in high-tech markets was further reinforced by the idea that uncompetitive markets often "correct themselves" and therefore that government intervention is unnecessary (Shelanski and Sidak 2001). The idea is that rapid technological changes favor markets where firms compete for temporary market dominance by innovating, only to be displaced by the next wave of product advancements. This process, famously described by Schumpeter as "creative destruction," and internalized by antitrust authorities as "competition for the market" (Evans and Schmalensee 2001) ensures that monopoly power is often fleeting (Schumpeter 1942: 81–90).[145] Moreover, "market processes

[145] Arrow (1962) complements Schumpeter by explaining in technical detail why potential entrants have strong incentives to innovate and therefore outcompete incumbent monopolists.

tend to protect consumers from the adverse effects of anticompetitive conduct, and these processes tend to work rapidly" (Lopatka 2000: 154) as one paradigm displaces the other and renders obsolete the exclusionary tactics that were once anticompetitive (see also Demsetz 1973).[146] Finally, how this fast-paced churn actually works in practice is usually inscrutable to outside observers, problematizing antitrust decision-makers' ability to regulate innovative firms effectively (Wright 2011: 230) and engendering unforeseen consequences.

Consequently, enforcement agencies took notice of this emerging consensus towards conservatism. The FTC's 2003 report, "To Promote Innovation: The Proper Balance of Competition and Patent Law and Policy," suggested that overly aggressive actions against firms

---

Arrow argues that a drastic innovation will shift the demand or supply curve out significantly, yielding an untapped source of rents for the innovator. Unlike an incumbent monopolist, a new entrant can capture these rents without cannibalizing existing profits, providing a stronger incentive for innovation. In this way, continuous competition for the market serves as a powerful driver of technological progress. This Arrowian idea was ratified by the DoJ when it acknowledged that "innovations in technology have led to competition from new entrants"—even in industries with high fixed costs, economies of scale and scope, and network effects, such as telecommunications (Delrahim 2018: 20).

[146] Moreover, according to Lopatka (2000: 154): "Antitrust enforcement is socially costly not simply because of false positives. Litigation itself is costly, both in the consumption of resources specialized to the judicial process and in its inevitable tendency to divert the attention of affected business personnel from productive endeavors. The costs of antitrust are minimized when claims that should fail are never brought."

engaging in legitimate patent licensing arrangements or strategic alliances could inadvertently stifle innovation by creating uncertainty and discouraging investment in R&D (see FTC 2003). In 2008, the DoJ's report, "Competition and Monopoly: Single-Firm Conduct Under Section 2 of the Sherman Act," underscored the importance of avoiding false positives qua excessive antitrust actions that might undermine incentives to innovate (see DoJ 2008), noting that certain business practices that appeared exclusionary in traditional markets could foster innovation and benefit consumers in rapidly evolving sectors like software and telecommunications (DoJ 2008). [147] As FTC Commissioner Maureen Ohlhausen later noted "[t]echnology industries are notoriously fast-paced, particularly industries involving the Internet. Poor or misguided antitrust enforcement action in such industries can have detrimental and long-lasting effects" (Ohlhausen 2013: 1).

Similarly, policymakers allowed several merger cases to sail through during the CDP era, particularly in sectors characterized by continual technological progress. For instance, the DoJ's 2018 challenge to the merger between AT&T and Time Warner was overruled on the grounds that it could stimulate innovation in content delivery and digital advertising (see AT&T 2018; AT&T 2019). Similarly, in 2007, the FTC allowed Google's acquisition of DoubleClick to proceed, reasoning that the merger could lead to significant advancements in online advertising technology that could benefit both advertisers and consumers (see FTC 2007). The US District Court for the Southern District of New York court approved the 2020 merger between T-Mobile and Sprint, highlighting the potential for an acceleration in the development and deployment of

[147] While this report was withdrawn by the DoJ in 2009 under President Obama, whose administration viewed it as too lenient (see DoJ 2009), its original publication remains the high-water mark of the Creative Destruction Paradigm's influence on single-firm conduct guidelines.

5G wireless technology (see Deutsche Telekom 2020). Additionally, the FTC cleared Amazon's 2017 acquisition of Whole Foods, arguing that the merger could drive innovation in the grocery retail sector that might improve logistics, delivery services, and store experiences (FTC 2017a).[148]

The more hands-off approach that characterized the CDP era marked a pendulum swing from cases such as *IBM* (1969), where the DoJ had accused IBM of monopolizing the computer market through "untoward practices" such as bundling its software with its mainframe hardware, exclusive dealing, and predatory pricing. The decade-plus litigation, which ended in dismissal in 1982, revealed how difficult it was for regulators to prove anticompetitive behavior in an industry marked by constant innovation and complex product integration (see IBM 1982; DeLamarter 1986). Enforcement agencies later cited this case as a cautionary tale: an example of a "false positive" that ended up harming consumer welfare by stifling innovation (see Crandall and Winston 2003; Levy and Weizer 1985).

[148] I hasten to emphasize, however, that the opposite was sometimes true. For example, the DoJ's challenge of the merger between General Motors and ZF Friedrichshafen in the early 1990s was partially based on the concern that it would consolidate the firms' R&D capabilities, thereby reducing competition in the development of new technologies, exemplifies this shift (see Kwoka and White 2014). By the same token, the 2010 Horizontal Merger Guidelines (see DoJ and FTC 2010) dedicated an entire section to fleshing out the concern that some mergers could stifle "innovation and product variety" by reducing the merged firm's incentives to innovate (ibid: Section 6.4).

As antitrust law evolved to accommodate the Schumpeterian view of innovation and dynamic competition, it faced a new challenge in the form of digital platforms, however. These platforms exhibited unique economic characteristics that I introduced at the beginning of this chapter and will review again below. The application of the Rule of Reason to these multisided markets was the next frontier in the ongoing adaptation of antitrust law to disruptive technological change.

THE RULE OF REASON APPLIED TO MULTISIDED PLATFORMS

The rapid ascendance of digital platforms reflects a series of changes to antitrust law. Multisided platforms, characterized by their ability to connect different groups of users, challenged traditional antitrust frameworks. The evolution towards tolerance—if not encouragement—of network effects and business practices employed by firms that operate digital platforms, which might have previously been frowned upon as exclusionary, marks a pivotal shift in antitrust enforcement and judicial interpretation (Rysman 2009).

This section examines the key court cases and enforcement agency regulations that incentivized mergers and business practices that culminated in today's Big Tech firms. To do so, I will do several things. First, define multisided markets and network effects. Second, explore the challenges antitrust decision-makers faced in defining relevant markets and how they settled on analyzing all sides and their interactions. Third, examine the increasing tolerance towards tying and diminished concerns over predatory pricing. Fourth, discuss antitrust decision-makers' growing acceptance of mergers that have allowed platforms to reach huge scales and develop enclosed ecosystems composed of myriad complementary services.

Understanding Network Effects in Multisided Markets

As previously discussed, multisided platforms operate by bringing together different user groups, such as consumers, advertisers, and developers, in a way that the participation of one group enhances the value of the platform to other groups. These network effects are central to the functioning of platforms like Facebook, Google, and Amazon. There are two types of network effects, direct and indirect. For the first type, the value of the platform increases directly with the number of users. For example, the more people who use a social media platform like Facebook, the more valuable it becomes to each user (Katz and Shapiro 1994). In terms of indirect network effects, the value of the platform increases for one group of users when another group grows. For instance, more consumers using Amazon attract more sellers, which in turn increases the platform's value to the consumers (Parker and Van Alstyne 2005).

Courts and enforcement agencies gradually developed an understanding of these dynamics, recognizing that multisided platforms often require different pricing strategies for different user groups. A platform might subsidize one side (e.g., offering free services to consumers) to attract users, while charging higher fees to another side (e.g., advertisers) that benefits from the large user base (Rochet and Tirole 2003).

However, antitrust decision-makers long faced difficulties in defining the relevant market when dealing with multisided platforms, let alone in assessing market power and competitive harm when one side of the market may not pay for the services they consume. The traditional approaches to market definition that I outlined above, such as the HMT, proved inadequate for the complex interactions and network effects exhibited by multisided platforms (Rysman 2009).

First, an orthodox antitrust analysis focused on only one side of the market could underestimate the platform's market power. For example, focusing only on consumers who use a platform for free might ignore the substantial market power the platform exerts over advertisers

who pay to access these consumers. Advertisers may depend heavily on reaching these users, giving the platform leverage to impose onerous terms and prices (Rochet and Tirole 2006).

Second, such an analysis might overlook the cross-market subsidization strategies central to platform business models (Armstrong 2006). Platforms build large user bases by offering free or low-cost services to one side of the market, thereby creating an attractive environment for participants on the other side willing to pay a premium for access (Rysman 2009). Once established, this dynamic enables exclusionary conduct: platforms can leverage user data to enhance their own products, negotiate exclusive arrangements, or integrate their offerings in ways that steer users away from rivals (Crémer et al. 2019).

Lastly, by not considering the interdependencies between different user groups, an orthodox antitrust analysis may fail to capture the full extent of competitive harm or benefits (Katz and Shapiro 1994). The credit card platform involved in *American Express* (2018) illustrates this point: the value of the platform to merchants depended on the number of cardholders, while the value to cardholders depended on merchant acceptance (Wright 2012). This interdependence meant that higher fees on merchants could reduce merchant participation, which would in turn diminish cardholder value and overall demand (Weyl 2010).

In light of these complexities, a key challenge for antitrust decision-makers has been devising ways to properly account for the value created by interactions between all sides of the market (Rysman 2009). Understanding how authorities grappled with these dynamics requires examining the key cases that shaped contemporary approaches to Big Tech companies like Google and Facebook.

Antitrust Decision-Makers' Early Struggles Understanding Platforms

The landmark Microsoft (2001) case was pivotal in shaping the legal *landscape* for digital platforms, highlighting the early challenges antitrust decision-makers faced when dealing with these emerging business models. The DoJ argued that by bundling the Internet Explorer browser with its dominant Windows operating system (OS), Microsoft stifled competition by discouraging users from choosing alternative web browsers such as Netscape Navigator. The court agreed, finding that this strategy allowed Microsoft to unlawfully maintain its monopoly in the personal computer OS market.

While the court recognized some aspects of the market's multisided nature, particularly the "applications barrier to entry" on the developer side (see Page and Lopatka 2007), it did not fully explore the role of developers or hardware manufacturers in sustaining Windows' dominance (Page and Lopatka 2007; Schmalensee 2002). The court's analysis primarily focused on the OS market for end users and the competitive effects of bundling within that context (Rubinfeld 2001).

This myopic focus on only one side of the market was a missed opportunity. On the one hand, it underappreciated Windows' network effects—how value for consumers depended on developer and hardware manufacturer participation—which were fundamental to Microsoft's market power (Evans et al. 2006). On the other hand, a more nuanced understanding of interdependencies could have led to a reassessment of Microsoft's bundling practice. Actions that seem anticompetitive on one side of the platform can produce pro-competitive effects on another (Rochet and Tirole 2003). For example, while the company's bundling of Explorer with Windows might have harmed competition in the web browser market, it could also have generated benefits across the platform—such as tighter OS-browser integration—that enhanced the overall user experience.

The Visa and MasterCard (2001) case represented a modest advance, with the court acknowledging the two-sided nature of credit card networks serving both consumers and merchants. Yet the court's focus remained primarily on the merchant side, particularly rules that barred member banks from issuing cards on rival networks like American Express and Discover. As a result, the market definition remained relatively traditional, without fully accounting for the intricate interdependencies between the different sides of the platform (Rochet and Tirole 2006).

FTC approval of Facebook's acquisition of Instagram in 2012 illustrated similar limitations. Focused on the free social media services enjoyed by users, and overlooking the complex interactions among the platform's users, advertisers, and developers, it allowed the merger to sail through without fully considering the potential for competitive harm over the long run, particularly Facebook's ability to leverage its dominant position across different sides of the market (Khan 2019). By the same token, the FTC's narrow market definition likely ignored potential efficiencies, such as advertisers' ability to leverage the integration of user data across platforms to better target users and the creation of a more robust ecosystem that could lead to greater user engagement and satisfaction (see Hovenkamp 2018; Carlton 2007; Wright 2014).

Growing Sophistication in Grasping Multisided Markets

By the late 2000s, a more nuanced approach to defining relevant markets—one that accounts for all sides of a platform—began to take hold, signaling a shift in antitrust jurisprudence (Katz and Sallet 2018). This evolution reflected antitrust decision-makers' growing appreciation for the unique economic characteristics of platform markets (Crémer et al. 2019).

Google's acquisition of DoubleClick in 2007 illustrates this development. In approving the merger, the FTC acknowledged the multisided nature of the digital advertising market. Google had argued that integrating its search data with DoubleClick's display ad-serving

technology would improve advertiser targeting, streamline ad-buying processes, reduce transaction costs, and drive innovation in ad formats and performance measurement (Evans 2008; Sokol and Comerford 2020). Despite concerns that the merger might endow Google with excessive market power by combining its dominance in search with DoubleClick's strength in display advertising, the FTC concluded that these potential efficiencies and innovations outweighed the potential harm (FTC 2007).[149]

The evolution towards a more sophisticated approach to platform markets crystallized in antitrust authorities' formal recognition that the value of a platform to each user group is intertwined and that changes on one side of it can significantly impact the other sides. The 2010 Horizontal Merger Guidelines highlighted this complexity, noting that "markets involving multisided platforms… may require more sophisticated economic analysis to account for the interdependencies between different sides of the platform" (DOJ and FTC 2010: 13). Accordingly, antitrust decision-makers adapted traditional tools like the HMT to platform contexts (Filistrucchi et al. 2014). Agencies began to assess whether a firm could profitably

[149] The FTC did not always apply this multisided framework consistently, however. When Facebook acquired Instagram in 2012, the agency focused narrowly on the free social media services enjoyed by users, overlooking the complex interactions among users, advertisers, and developers. While the merger gained approval without full consideration of Facebook's ability to leverage its dominant position across different sides of the market (Khan 2019; Wu 2018), the FTC's narrow market definition likely ignored potential efficiencies, such as improved advertiser targeting through integrated user data (Hovenkamp 2018; Carlton 2007; Wright 2014).

impose a SSNIP on one side of the platform without losing too many customers on the other side (Boudreau and Hagiu 2009).[150]

These developments set the stage for a pivotal moment in antitrust law, American Express (2018), where the Supreme Court fully recognized the complexities of multisided platforms and the necessity of considering interdependencies between different sides of the market (Evans and Schmalensee 2016; Rochet and Tirole 2006). The United States and several states sued American Express (Amex), arguing that its anti-steering provisions—which prevented merchants from directing customers to use lower-cost credit cards—were anticompetitive: they effectively maintained and enhanced Amex's market power, resulting in higher fees for merchants, which were then passed onto consumers (Wright and Yun 2018).

[150] To define platform markets more accurately, antitrust agencies began to rely on advanced economic models. Discrete Choice Models (DCMs) allow analysts to estimate how consumers perceive and prioritize different product features, including price, quality, and network effects, enabling evaluation of how changes in these features might influence consumer decisions across different sides of a platform. For instance, a DCM can quantify the extent to which a price increase for advertisers might impact user engagement (Train 2009). Conjoint Analysis (CA) proved similarly useful, helping analysts infer consumer preferences by presenting respondents with product or service combinations and asking them to rank or choose between them. Armed with these insights, CA can simulate consumer substitution patterns across a platform, revealing how consumers might switch platforms or adjust usage in response to changes in pricing or service quality (Louviere et al. 2000).

The key issue was the definition of the relevant market. Amex argued that merchant and cardholder services were interdependent and should be treated as a single market (Evans and Schmalensee 2016). It maintained that its anti-steering provisions were essential for sustaining premium services—rewards programs, travel perks, and superior customer service—by ensuring the necessary revenue (Wright and Yun 2018). These provisions allowed Amex to charge higher merchant fees, which funded innovations in payment technology and security alongside enhanced cardholder benefits (Rochet and Tirole 2006). By preventing merchants from steering customers to lower-cost cards, Amex argued, it preserved the balance between the two sides of the platform necessary to deliver these benefits.

The Supreme Court largely agreed, holding that both the merchant and consumer sides of the credit card platform must be considered together when defining the relevant market. Because the plaintiffs had failed to show anticompetitive effects in the correctly defined market—which included both merchants and cardholders—evidence of increased merchant fees alone was insufficient to establish harm without considering the impact on cardholders.

Unlike the courts, however, enforcement agencies did not necessarily have a "Come to Jesus Moment" about multisided markets. Consider the DoJ's attempt to block Sabre's acquisition of Farelogix in 2019 (DoJ 2019; DoJ 2020). Sabre operated the largest global distribution system (GDS) in the United States—a digital platform connecting airlines with travel agencies that handles over half of all airline bookings through travel agents. Farelogix, by contrast, had developed "new distribution capability" technology enabling airlines to bypass traditional GDS platforms and offer more personalized fare bundles directly to travelers booking through agencies. The DoJ characterized the $360 million acquisition as a dominant firm's attempt to eliminate a disruptive competitor. Yet in defining the relevant market narrowly—

framed around airline booking services sold through travel agencies—the DoJ's litigation theory treated the platform largely as a one-sided booking-services problem rather than as a platform requiring a two-sided accounting of cross-group effects (DoJ 2020).

Echoing the logic that prevailed in American Express (2018) regarding two-sided transaction platforms, Sabre argued that the competitive assessment should instead reflect the platform's multisided structure and related network effects (American Express 2018; Sabre 2020). As in the American Express case, the court ultimately ruled for the defendant, agreeing with its logic and concluding that the DoJ had not shown that the acquisition was likely to substantially lessen competition (Sabre 2020).[151]

Increased Tolerance Towards Tying and Predatory Pricing

Along with a more nuanced approach to platforms, antitrust decision-makers also exercised growing leniency towards business practices that would have previously been deemed exclusionary. The Rule of Reason approach increasingly acknowledged that such practices, while potentially exclusionary in certain contexts, could also enhance consumer welfare by fostering innovation, increasing convenience, and improving product quality.

While tying arrangements were historically viewed with suspicion under antitrust law, as they were seen as tools for leveraging market power from one product to another (Bork 1978; Hovenkamp 2005), the Supreme Court's decision in *Jefferson Parish Hospital District No. 2*

[151] While the court did not explicitly criticize the DoJ's narrow market definition, its decision implied that focusing solely on the single-sided market for airline booking services may have obscured the platform's full scope of competitive interactions and potential effects across all user groups (see Rysman and Wright 2014).

(1984) marked a pivotal shift. While the Court found no illegal tie in this instance—because the hospital lacked sufficient market power over patients in the relevant market—it opened the door for a more nuanced approach by acknowledging that tying could sometimes be justified by legitimate, procompetitive reasons. These justifications might include cost savings, where tying allows for efficiencies in production or distribution, or quality control, where a bundled offering ensures a higher standard of service or product performance (see Sullivan and Grimes 2006).

Simultaneously, antitrust decision-makers exhibited greater tolerance of aggressive pricing strategies in competitive markets. A seminal 1975 article by Areeda and Turner laid the intellectual foundation for this shift by proposing a cost-based test for predatory pricing, arguing that prices should only be considered predatory if they fall below an appropriate measure of cost—specifically, average variable cost as a proxy for marginal cost (Areeda and Turner 1975). US Court of Appeals Judge Stephen Breyer notably adopted this approach in Barry Wright (1983), emphasizing that, because "false positives" in predatory pricing cases could stifle legitimate competitive behavior, it was eminently important to rely on sound economic analysis when distinguishing predatory conduct from vigorous price competition.[152]

This doctrinal shift had profound implications for the "Superstar" firms I discussed in Chapter 4. By establishing that low prices were lawful so long as they remained above the dominant firm's own measure of cost, the courts effectively created a safe harbor for Limit

[152] The Supreme Court reinforced and extended this framework in Matsushita Electric Industrial (1986) and Brooke Group (1993), adding the requirement that plaintiffs demonstrate a dangerous probability that the alleged predator could recoup its losses.

Pricing (see Edlin 2002), which I explained in technical detail there.[153] A dominant firm could now strategically set its prices just below the costs of its less efficient rivals—thereby excluding them from the market—without triggering antitrust liability. So long as the incumbent was more efficient than the entrant, it could use that efficiency to maintain a monopoly position indefinitely.

Applying this Permissiveness to Platforms

Evolving attitudes towards tying and predatory pricing were particularly salient to the ongoing dialectic between the economics of multisided markets and antitrust. Antitrust decision-makers grew to appreciate the fact that these practices can strengthen network effects and enable platforms to utilize resources from one side of the market to drive innovation and create value on the other (Rochet and Tirole 2006)—especially when data from one group (e.g., consumers) can be leveraged to benefit another group (e.g., advertisers or developers). DoJ officials went so far as to emphasize that firms seeking to create network effects with both direct and indirect benefits to users may find it rational to price very aggressively—even giving products away—to build an

[153] To recapitulate, the Superstar acts as a price maker but acts as if it faces competition: it engages in Limit Pricing by setting its price just below the marginal cost of the inefficient "competitive fringe" (MC_{fringe}) to deter entry. Consequently, while the firm may superficially resemble a monopolist in terms of scale, its excess profits are technically Ricardian Rents—derived from its superior efficiency relative to the fringe ($MC_{fringe} - MC_{Star}$)—rather than Monopoly Rents derived from restricting output to raise prices above the competitive ceiling.

installed base, and that such dynamic pricing can be consistent with vigorous competition (see Klein 1998; Shapiro 1996).

Platforms responded in kind, exploiting the more permissive antitrust environment to gain market share and increase profits via innovative business practices that drove increased user engagement. With respect to tying, Apple required that all iOS apps be distributed through its App Store, accompanied by a 30% commission on in-app purchases. This policy compelled developers to use Apple's payment system if they wanted to reach iOS users, effectively tying the App Store's distribution service to its payment processing service—although the company has faced subsequent legal challenges to this model that I will discuss in Chapter 12. Similarly, Google required Android device manufacturers to pre-install Google Search and Chrome as a condition for accessing its essential Google Play Store. With respect to aggressive pricing, Amazon justified its strategies on the grounds that they deliver significant consumer benefits such as lower prices and wider product availability (see Khan 2017).[154]

Permissiveness Towards Mergers

As antitrust guidelines evolved to address the complexities of multisided platforms, a more permissive stance towards mergers in general also emerged, particularly in high-technology markets. This leniency allowed tech giants to strategically fold in other firms operating in complementary businesses and scale up. In turn, it enabled them to build expansive, integrated

[154] For enforcement agencies to succeed in a predatory pricing claim against Amazon, they would have to prove that its pricing behavior meets the stringent requirements established in Brooke Group (1993): both below-cost pricing and a dangerous probability of recoupment.

ecosystems that leveraged network effects, data integration, and R&D across multiple market segments.

The foundation was laid by several iterations of the DoJ's Merger Guidelines, which starting in 1982 prioritized consumer welfare and economic efficiency over rigid concerns about market structure and the number of competitors (DoJ 1982; DoJ and FTC 1992; DoJ and FTC 2010). This shift allowed regulators to consider whether a merger would generate efficiencies—such as cost savings, improved production, or enhanced product offerings—that could ultimately benefit consumers (Carlton and Perloff 2005). By 1992, the guidelines explicitly recognized that mergers could foster innovation and drive competition even in concentrated markets (Kwoka and White 2014). As that vintage noted, "efficiencies can enhance the ability and incentive of the merged firm to compete, which may result in lower prices, improved quality, or innovation" (DoJ and FTC 1992: 30). This permissive stance was particularly influential in the tech sector, where economies of scale and rapid innovation are crucial competitive advantages (Schmalensee 2000). Finally, the 2010 revisions marked a pivotal shift in antitrust enforcement, particularly for high-tech and digital markets (DoJ and FTC 2010), as this vintage placed a greater emphasis on innovation, recognizing the importance of network effects and complementary assets for the first time (see Shapiro 2010b).

These developments were paralleled by a similar evolution toward greater tolerance of vertical integration.[155] The intellectual foundation for this permissiveness was laid in the 1970s

[155] Before the shift toward permissiveness, certain forms of vertical integration met strong resistance from antitrust decision-makers. This was evident in the case of A&P, a grocery chain that elicited aggressive enforcement actions—it was sued under three statutes: the Sherman Act,

by economists like Oliver Williamson, who argued that vertical integration could reduce transaction costs—expenses related to negotiating, monitoring, and enforcing contracts in the marketplace—and improve efficiency (Williamson 1985).[156] This shift in thinking was reflected in legal decisions such as *Continental T.V.* (1977), where the Supreme Court recognized the procompetitive potential of vertical restraints, setting a precedent for a more lenient approach to vertical acquisitions (see Pitofsky 2001). It was further reinforced by the 1984 Non-Horizontal Merger Guidelines, which explicitly acknowledged that vertical mergers could create efficiencies by improving coordination across different stages of the supply chain (DoJ 1984: 5-6). Finally, in 2018 a federal court allowed AT&T to acquire Time Warner, accepting the former's argument that the merger would promote efficiency and innovation via the integration of content creation with distribution (AT&T 2018).

On the back of increased tolerance for all manner of previously verboten unilateral firm behavior and more permissive horizontal and vertical merger review, digital platforms grew into

---

Clayton Act, and Robinson-Patman Act—in the mid-20th century after it developed a highly integrated industrial structure, controlling everything from manufacturing to retailing (see A&P 1949). These cases precipitated fines and orders that A&P end both pricing policies and business practices, such as exclusive deals forged with suppliers, that "unfairly undercut" its competitors (see Levin 2011).

[156] Williamson argued that this is particularly important in industries with complex, asset-specific investments that create a dependency between transaction partners and exposure to opportunism. Vertical integration may also reduce double marginalization and thus increase production and eliminate deadweight losses.

unprecedentedly large and complex ecosystems. These ecosystems eventually evolved into the AI hyperscalers that, as Chapter 10 will show, spent hundreds of billions of dollars training AI models and building data centers or expanding cloud-computing capacity to run AI inferences.

Mergers and Big Tech Empire-Building

The permissive merger regime I outlined above reached its apogee with Big Tech's acquisition spree during the early aughts and 2010s, when antitrust agencies greenlit the deals that stitched together complementary data ecosystems. Far from mere consolidation, these mergers created cross-market data reservoirs—search + video, social + messaging, retail + cloud.

Consider several examples. Facebook's acquisition of Instagram in 2012 and WhatsApp in 2014 consolidated its position in social media by absorbing potential rivals. Google secured dominance in online advertising by integrating DoubleClick's ad-serving technology with its search engine in 2007, while its 2006 acquisition of YouTube gave it control over the rapidly growing online video market. Amazon's acquisition of Whole Foods in 2017 merged its online platform with physical retail, strengthening its distribution network and generating consumer insights to support targeted advertising.

At the same time, the FTC exercised leniency towards Google's business practices, choosing not to pursue action against the company for putatively exclusionary practices. In 2013, the agency closed its inquiry into Google's search practices after investigating allegations that the platform manipulated its algorithms to favor its own services over competitors (FTC 2013b). The FTC also declined to act against (i) Google's AdSense exclusivity agreements, which prevented websites from displaying ads from rival networks; (ii) its scraping of content from other websites to display on its own search results; and (iii) its restrictions on advertisers' ability to port campaigns to competing platforms (ibid). Instead, the agency asked Google to carve out content

opt-outs—allow websites to exclude their content from Google's specialized search results without being delisted from organic search—and improve advertiser portability.[157]

In the shadow of antitrust agencies' chronic forbearance, Google refined business practices that cemented its dominance across multiple digital markets.[158] By creating synergies between internet search, mobile operating systems, online video, and digital advertising, Google set off a flywheel that transformed user data into personalized experiences, boosting engagement and commanding premium advertising rates, which funded further R&D and strategic acquisitions that expanded the ecosystem and, ultimately, supplied the training data for its AI models.[159] I flesh out that phenomenon in the next chapter, along with recounting how Facebook and Apple also achieved digital hegemony in a similar fashion.

---

[157] Critics argued that this decision was wrongheaded because Google's practices indeed stifled competition (Hauck 2015); see the next paragraph.

[158] For example, in 2016 Google began requiring advertisers to use its tools to buy ads, rather than tools offered by other companies, on its subsidiary YouTube, which at the time was really the only game in town for posting and viewing online videos.

[159] Of course, this forbearance did not last indefinitely: In August 2024, Google lost a significant antitrust case against the DOJ. The ruling found that Google had illegally monopolized online search and advertising by paying companies like Apple and Samsung to install Google as the default search engine on smartphones and web browsers. The court deemed that this practice stifled competition and harmed consumers. In Chapter 12 I briefly discuss some antitrust cases that have upended the CDP-type approach to antitrust explored in this chapter.

To preview that story here: Facebook (later known as Meta) leveraged its control over Facebook, Facebook Messenger, WhatsApp, and Instagram to amass unparalleled social graph data, monetize attention through hyper-targeted advertising, and reinvest those profits into acquisitions that preempted competitive threats while generating still more data to refine its algorithms; Apple's walled garden approach to the iPhone, Apple Watch and iPad allowed it to capture premium consumers within a tightly integrated ecosystem, extract rents through App Store commissions and services subscriptions, and leverage that revenue to fund hardware innovation that deepened user lock-in.

By the early 2020s, however, the initial leniency that facilitated the rise of these Big Tech firms was met with growing skepticism: regulators and courts increased their scrutiny of mergers and platform strategies that were once seen as beneficial. Earlier warnings that Google might leverage DoubleClick's dominance in display advertising to foreclose competitors (Evans 2008), or that Facebook engaged in tying and predatory pricing to stifle competition (Wu 2018), now appeared prescient. The bigger concern was that Big Tech firms had become too big to rein in and might translate their economic clout into political power, eroding democracy in the process (Wu 2018). This growing "techlash" signaled a shift away from the permissive antitrust environment that allowed these ecosystems to flourish and potentially marked a return to the more populist approach to antitrust that I outlined at the start of this chapter, a topic I will address in Chapter 12.

CONCLUSION

This chapter has traced the century-long evolution of American antitrust from a populist crusade against "bigness" to a technocratic exercise in consumer welfare analysis. The Brandeisian distrust of concentrated economic power that dominated the mid-twentieth century gave way, not

through ideological conversion to Chicago School orthodoxy, but through the organic necessities of the Rule of Reason itself. Faced with real cases requiring proof of competitive harm, courts and enforcement agencies turned to the machinery of economics: price theory to define relevant markets and measure market power, the Hypothetical Monopolist Test to avoid the Cellophane Fallacy, and Schumpeterian insights to appreciate the role of temporary monopoly rents in spurring innovation. The resulting doctrinal shifts—from per se prohibitions to case-by-case balancing, from hostility toward vertical integration to recognition of its efficiency benefits, from reflexive condemnation of tying and aggressive pricing to safe harbors for conduct above cost—created the permissive legal oxygen that allowed digital platforms to achieve unprecedented scale.

This evolution culminated in a sophisticated, if contested, framework for analyzing multisided markets. From the early struggles in Microsoft (2001) to the Supreme Court's recognition in American Express (2018) that both sides of a platform must be considered together, antitrust decision-makers gradually internalized the economics of network effects, cross-subsidization, and platform interdependencies. The Merger Guidelines progressively emphasized efficiency and innovation over structural concerns, greenlighting the acquisitions that stitched together the data ecosystems of Google, Facebook, and Amazon. It is doubtful that smaller, fragmented firms without these integrated platforms could have generated the vast, diverse, real-time data pools and constructed the concomitant data refineries required to train LLMs.

Yet by the early 2020s, the initial leniency that facilitated these empires was met with growing skepticism—a "techlash" reflecting concerns that Big Tech had become too powerful to discipline and might translate economic dominance into political power. This signaled a potential

return to the populist sensibilities that opened this chapter, a topic I will address in Chapter 12. In the next chapter, I explain how, together with permissive antitrust policies, the 1996 Telecommunications Act, Section 230's liability shield, and a "contract-over-property" framework that vested data ownership in extractors rather than users enabled platforms to build the data flywheels that ultimately fueled the AI revolution.

CHAPTER 8: DIGITAL PLATFORMS, DATA FLYWHEELS

What explains the AI revolution? In most proximate terms, it was a Cambrian-like explosion of data, which modelers used to train LLMs, leading many pundits to declare that data is the new oil of the AI age (e.g., The Economist 2017). To get their models to grasp and generate humanlike language, engineers feed them billions of geo-tagged social media posts, search queries, purchase behaviors, product reviews, blog comments, memes, marketplace listings, collaboratively written encyclopedia entries, and software code repositories.

Like oil, data is abundant and cheap in its raw form, but only economically valuable after its refined (The Economist 2020).[160] To convert crushing amounts of data into useful insights that can be readily bought and sold, digital platforms must put in the work: invest in data curation, algorithmic processing, display ranking—deciding where content appears on the webpage or feed—and filtering, which determines whether an item is eligible in the first place (Parker et al. 2016). This creates an essential "barter deal" (O'Reilly 2021): platforms provide multibillion dollar services like global search and real-time mapping in exchange for individual data points that, on their own, are worth mere pennies.

Like oil refiners that own and control everything from the wellhead to the gas station, vertical integration helps these platforms' efforts to commercialize their product: the type of integrated digital ecosystems I discussed in the previous chapter ensure that the data flowing from one service (like search) can be instantly used to refine the performance of another (like

[160] In this vein, Tim O'Reilly (2021) argues that data is perhaps better understood as "the new sand"—ubiquitous and nearly worthless in its raw form, but economically valuable when refined into the silicon that goes into computer chips.

targeted ads). Digital platforms that capture data at every point of user contact and feed it back into their primary monetization engines set off a virtuous circle. Ever improving, more useful search results and increasingly relevant and entertaining content lock in user loyalty (Shapiro and Varian 1999) and serve more effective ads, attracting a larger pool of advertisers willing to pay top dollar to reach more eyeballs (Rochet and Tirole 2003). The resulting revenue stream then funds publisher partnerships and new product development, further amplifying user engagement and entrenching network effects (Shapiro and Varian 1999). This is all made possible by a seamless pipeline where user activity is captured, refined, and monetized within a single proprietary environment.

Looking Under the Hood

Continuing with this book's new institutional economics approach, in this chapter I'll explore how the federal government laid the foundation for the data economy, examining the political economy of that system—what makes data a unique resource that can be commercialized, exchanged and commodified—and fleshing out the data flywheel's industrial organization. I'll also outline the political and legal choices that enabled massive user participation online. Besides the permissive antitrust regime I documented in the previous chapter, this includes a "contract-over-property" framework that favors data extractors, the absence of a database right at the federal level, the enforcement of "clickwrap" licensing agreements and the liability protections of Section 230 of the 1996 Telecommunications Act.

Among these, the antitrust regime's permissiveness—which I documented in the previous chapter—created the diverse data reservoirs essential for AI. The shift from populist suspicion of bigness to a consumer welfare standard that tolerated "free" services and recognized the efficiency benefits of vertical integration allowed platforms to build precisely the cross-

subsidized, data-extractive business models described above. The "data reserves" required to power deep learning architectures are not just about volume, but about variety. By allowing tech giants to accumulate firms in adjacent markets—integrating maps with search, or social media with hardware—and by analyzing platforms as multisided markets where value on one side depends on participation on the other, antitrust authorities facilitated the creation of a "data commons" that was deep enough to move AI from simple pattern matching to sophisticated reasoning. By focusing on competitive effects rather than market share, they greenlit Big Tech firms to justify mergers that expanded their ecosystems by integrating new technologies, talent, and customer bases (Shelanski 2011). As a result, Big Tech firms pursued acquisitions more aggressively, further entrenching their market dominance and enabling them to leverage network effects to an even greater extent (Baker and Shapiro 2010).

This regulatory green light was the necessary condition for "virtuous cycles" of innovation, diffusion, and more innovation. By permitting the seamless orchestration of hardware, software, and services (often via vertical restraints rather than ownership), antitrust authorities allowed firms to coordinate complex, disintegrated supply chains into unified user experiences. Had the courts adhered to the older, "structuralist" suspicion of tying and vertical control, the tightly integrated PC and smartphone ecosystems that drove the Third Industrial Revolution (see Chapters 5 and 6) might have been fragmented by regulation, potentially stalling the very diffusion process that democratized computing. These antitrust-enabled data reservoirs, in turn, became the raw material that contract and copyright frameworks then refined into proprietary assets.

Permissive copyright interpretations, alongside the contract frameworks above, established a "scrape first, ask permission later" culture, treating the web's billions of user-

produced pages as freely harvestable resources ripe for the taking—albeit changes in digital platforms' business models and user engagement strategies eventually led to a less permissive approach to data tracking and profile making over time.

Meanwhile, later in the chapter, I will discuss how digital platforms—led by Google's pioneering digital terraforming—perfected user engagement practices that opened the data spigot, converting billions of disparate human interactions into the structured datasets that ultimately advanced the AI revolution.

This chapter proceeds as follows. The first section traces the telecommunications policies and federal investments that built the physical and regulatory infrastructure of the early internet. The second section examines the political economy of data, exploring its unique properties as a non-rivalrous but excludable good, the property rights regimes governing its collection and use, and the privacy regulations that increasingly constrain its exploitation. The third analyzes Section 230's liability shield and how it enabled platforms to scale user-generated content without crippling legal risk, fostering the content-rich, personalized environments that generated massive training datasets. The fourth section traces the technologies, norms, and legal interpretations—from robots.txt protocols to fair use doctrine—that sustained the open web as a freely harvestable resource for AI developers. The remaining sections then engage three platform case studies: Google's integrated advertising flywheel and its contributions to search, recommendation, and natural language processing; Facebook's engagement-driven business model and its investments in computer vision and content recommendation; and Apple's privacy-centric approach to on-device AI. A conclusion synthesizes how these platform-dynamics positioned the United States to lead the AI revolution.

THE INFRASTRUCTURE OF CONNECTIVITY: FROM SILICON TO SPECTRUM

The previous chapter explored how, during the CDP era, US antitrust authorities greenlit digital platforms to grow into global behemoths that acquired network effects, both direct and indirect, and build interoperable ecosystems. Simultaneously, the federal government moved beyond its traditional role as a utility regulator to become an architect of digital infrastructure. This transition was codified in the Telecommunications Act of 1996, a landmark piece of legislation designed to replace the rigid monopolies of the landline era with a market-based framework for the digital age (US Congress 1996). While the Act fundamentally transformed the US telecommunications landscape into a more competitive and geographically expansive marketplace, it also helped usher in the Fourth Industrial Revolution by underwriting the buildout of cellular infrastructure that was critical to the digital economy and eventually the commercialization of AI.

Incentivizing Competition and Investment

The Act's most significant contribution to the digital revolution was its treatment of spectrum as a market-based resource.[161] Building on the success of an initial round of "simultaneous multiple-round" auctions in 1994, the federal government institutionalized the use of competitive bidding to allocate radio spectrum (Cramton 1997; Milgrom 2004). By 2008, several of these auctions had transferred vast swaths of high-quality "airway real estate" from television

[161] This represented a fundamental break from the philosophy of the Communications Act of 1934. Under the old regime, telecommunications infrastructure was treated as a "natural monopoly"—a market where high fixed costs made competition inefficient and untenable (see US Congress 1934; Laffont and Tirole 2000).

broadcasters to cellular providers, providing the "beachfront property" necessary for the 4G LTE buildout I discussed in Chapter 6 (see FCC 2008).

Between 2005 and 2008, the FCC significantly expanded available wireless airwaves by auctioning over 3,000 licenses across several new radio spectrum bands. Verizon and AT&T won the auctions over the A- and B-Blocks, radio wave bandwidth that became the backbone of their future 4G rollouts (Cramton 2010). The 700 MHz auction that followed was a watershed moment in Creative Destruction Paradigm policymaking because it coupled market-based allocation with "open platform" requirements. Unlike previous auctions, the FCC mandated that Verizon, the winner of the "C-Block" spectrum—a 22-megahertz band of high-quality airwaves formerly used for analog television—allow users to download any application and use any device of their choosing on the network, effectively extending the interoperability principles of the computer industry to the wireless carrier market (Crawford 2013; Lessig 2001). In doing so, federal policy ensured that the mobile internet remained a "generative" platform where innovation could happen at the edge, rather than being restricted by the owners of the underlying infrastructure (Zittrain 2008).

Through the 1996 Telecommunications Act and subsequent FCC implementations and judicial rulings, the federal government provided institutional certainty for firms to invest over longer horizons, incentivizing them to deploy massive capital outlays towards the internet's next generation wireless infrastructure. Let's consider a few of the most notable developments.

The 1996 Telecommunications Act also established a 90-day to 150-day window for local governments to process infrastructure applications (see FCC 2009). This "shot clock" rule preempted local "NIMBY" (Not In My Backyard) obstructionism by ensuring that municipal

zoning boards could not effectively ban new cell towers through indefinite delays, thereby accelerating the physical buildout of the wireless grid (see City of Arlington 2013).

By the same token, the Act relaxed ownership caps and largely abandoned unbundling mandates—requirements that dominant providers lease portions of their physical infrastructure to competitors at regulated rates—for high-speed broadband and fiber-optic networks (see FCC 2003; FCC 2005). This signaled that while the government might mandate competition on legacy copper wires, it would allow firms to reap the full rewards of risky capital improvements in new generations of telecommunications infrastructure.

Conversely, for legacy copper lines, the Act required incumbent carriers to lease portions of their networks to competitors at cost-based rates. This prevented incumbent carriers from denying new entrants access to the "last mile" of physical wire reaching the consumer (see Nuechterlein and Weiser 2005). While this was not a common carrier provision per se, it did share the fundamental DNA of utility-style regulation: treating the physical network as a public utility that must be accessible to all comers rather than as private property.[162]

Political Economy Dimensions of these Reforms

By making opportunism by local zoning authorities, future regulators, and rival firms next to impossible, the 1996 Telecommunications Act addressed the hold-up problem I explore in

---

[162] The Act's 1934 predecessor classified providers (most notably AT&T/The Bell System) as "common carriers", subjecting them to strict rate regulation and nondiscrimination mandates ("universal service"). In exchange, the state protected them from market forces by erecting high barriers to entry that rendered competition effectively illegal (see Vietor 1994; Temin and Galambos 1987).

various places in this book (see Chapters 3 and 10). Once a telecommunications firm makes a fixed, irreversible, and asset-specific investment—such as sinking billions into cellular towers and fiber-optic backhaul—those sunk costs cannot be redeployed to other uses. This creates a vulnerability where a firm can be "squeezed" by regulators or competitors that prevent it from offsetting its long-run costs (Klein et al. 1978; Williamson 1985; Galetovic and Haber 2017). The shift toward stable, market-based spectrum auctions meant that carriers could secure reliable, long-term property rights to the airwaves, protecting them from the vagaries of administrative reallocation and allowing them to borrow against these assets to finance network expansion. Easing cell-tower siting restrictions through the FCC "shot clock" rule meant that firms faced lower transaction costs and less uncertainty when physically deploying their networks, preventing local municipalities from holding capital investments hostage.

Moreover, relaxing ownership caps and largely abandoning unbundling mandates for next-generation networks meant that regulators credibly committed to not expropriating the ex-post value of network upgrades, ensuring that the returns on sunk costs would accrue to the investor rather than being dissipated by mandated sharing (see FCC 2005).[163] This regulatory shift therefore solved a critical coordination problem (Rodrik 2004; Hausmann and Rodrik 2003). If the government had kept "unbundling" rules for new fiber, Competitive Local

[163] This move was controversial and heavily litigated (see, e.g., Verizon 2002). While proponents argued that "managed competition" of this sort forced the transition from slow, circuit-switched telephone lines to modern high-speed broadband networks, critics countered that unbundling mandates actually suppressed investment by reducing incumbents' expected returns (Crandall 2001; Hausman and Sidak 1999; Pindyck 2004).

Exchange Carriers (CLECs), upstart telecom companies like MCI that piggybacked on Verizon or AT&T's copper wires, could have waited for those incumbents to spend billions digging trenches and laying down fiberoptic cable, only to then demand access to those lines—thereby forcing the latter to bear all the downside costs and risk while capping their upside potential (Pindyck 2007; Hausman 1997). By removing this "free rider" threat, the government restored the "first mover advantage" necessary to trigger the initial capital outlay.

Indeed, in the telecommunications sector, the "first mover advantage" is not merely a strategic edge but a structural economic reality. Building physical networks involves massive upfront sunk costs but very low marginal costs to serve each additional user. This cost structure creates a "natural monopoly" dynamic at the local level: once a firm has wired a neighborhood, it is economically irrational for a second competitor to replicate that infrastructure merely to split the customer base (see Laffont and Tirole 2000). Consequently, the first firm to deploy effectively captures the market, locking in subscribers through high switching costs. By eliminating unbundling mandates, the government allowed firms to rely on this winner-take-all prize as the justification for their initial risk (see Pindyck 2007).

The Macroeconomic Impact: From Infrastructure to Innovation

The various telecommunications reforms that were downstream from the 1996 Telecommunications Act had profound effects. They induced price and access convergence, transformed communications networks and cemented an open digital transmission frontier. They also set off an epic investment boom—one that foreshadowed the even bigger data center blowout of the early 2020s (see Chapter 11).

Within three years of the Act, close to 150 new Competitive Local Exchange Carriers (CLECs) entered the market, driving down (quality adjusted) internet connection costs by

roughly 20% (Greenstein and Stranger 2007). Increased affordability was matched by rapid infrastructure expansion; by 2000, 98% of US public schools boasted internet access, a feat largely attributed to the E-rate program and related pro-investment initiatives (NCES 2000).

Under the legacy monopoly model, incumbents were slow to deploy broadband technologies like ISDN or DSL. The threat of competition—specifically "intermodal competition" between cable and telephone companies—forced a sudden acceleration in network upgrades. As cable providers began offering voice services and phone companies entered video markets, the "bundled" service model emerged, providing the reliable bandwidth necessary for bandwidth-intensive platforms like YouTube and Netflix to thrive (Goldfarb 2007; FCC 2013).

Perhaps the quietest, yet most consequential reform was the "Open Access" requirement of the 2008 700 MHz C-Block auction. By mandating that network owners allow users to use any device or application of their choosing, the FCC established a regulatory precedent for "device-level neutrality." This sent a clear signal to Silicon Valley: the cellular network would not be a "walled garden" controlled by carriers, but a vibrant, open ecosystem. This ethos directly enabled the smartphone-driven digital economy, ensuring that as 4G LTE rolled out, the "innovation at the edge" could outpace the owners of the underlying spectrum (Zittrain 2008; FCC 2014).

The 1996 Telecommunications Act and its subsequent FCC implementations did more than just lower phone bills; they fundamentally reconfigured the American economy. By replacing utility-style "command and control" with market-based incentives, the reforms unleashed approximately $1.4 trillion in private sector investment over the following decade (Markey and Fields 2016). It was upon this infrastructure that the next generation of digital

platforms would build their empires and eventually help launch the AI-driven world we inhabit today.

WORLDCLASS PIPES: NECESSARY BUT NOT SUFFICIENT

While versatile and powerful AI required a telecommunications superhighway, model developers required nimble and agile vehicles to traverse it: dynamic, context-rich human language generated on GPS-enabled mobile devices. To be sure, foundational models relied on large-scale web crawls to mine digital libraries, academic papers, and news archives. However, these traditional sources alone were ultimately inadequate: Geocoded, mobile-first user-generated data captured both mass trends and long-tail nuances—obscure facts, specialized vocabulary, and unique cultural expressions—that provided models with a level of situational awareness and linguistic fluidity that static datasets could not replicate.

The immense value of this real-time, conversational data for training AI was starkly illustrated by the drastic measures some platforms took to protect it. A prime example was Elon Musk's controversial decision in July 2023 to impose strict "rate limits" on Twitter (which he later renamed X), temporarily restricting the number of posts users could read each day (Kovach 2023). Musk justified the move as a necessary defense against “extreme levels of data scraping” by AI companies, which he argued was degrading the service for regular users (Musk 2023).

While Musk’s restrictive policy caused widespread frustration (see Reuters 2023), it revealed the possibility of a fundamental shift in the digital ecosystem. The vast, freely accessible stream of human conversation that had been a taken-for-granted by AI developers was

now being declared as a highly valuable, finite asset to be controlled and monetized.[164] The volatile CEO's willingness to risk alienating his user base by building a wall around his platform's data underscored just how vital the digital commons had become in the race to build the next generation of AI.

THE POLITICAL ECONOMY OF DATA

Data exhibits unique economic properties that distinguish it from traditional goods. It is fundamentally non-rivalrous: once created, it can be copied and used by multiple parties simultaneously without depletion, allowing infinite reuse in applications like machine learning algorithms (see Chapter 2 of this book). However, it is often excludable, as access can be restricted via encryption or paywalls, positioning it as a club good—available to select groups but not universally public (see The Economist 2020; Chapter 2 of this book).

When exploited by platforms, data also enjoys economies of scale and scope. While more data begets better analysis and predictions, "better" data, in terms of its quality, diversity, and recency, does so as well (McAfee and Brynjolfsson 2012). Take ridesharing as an example:

> By itself, an individual ride on Uber is high value for both rider and driver—a desirable core interaction. As the number of platform participants increases, so does the value Uber

[164] To be sure, not every platform freely shared its data at that point; Facebook, for example, largely walled off their user data, and legal/privacy developments had restricted some scraping practices. But, enough of the user-generated canon was publicly accessible during the late 2010s and early 2020s, giving LLMs no shortage of examples to learn from.

> delivers to both sides of the market; it becomes easier for consumers to get rides and for drivers to find fares. Spillover effects further enhance the value of Uber to participants: Data from riders' interactions with drivers—ratings of drivers and riders—improves the value of the platform to other users. Similarly, data on how well a given ride matched a rider's needs helps determine optimal pricing across the platform—another important spillover effect (Alstyne et al. 2016).

In terms of the ease of entry by new players into data centric markets, it's a mixed bag. On the one hand, innovations in data storage and processing afforded by cloud computing profoundly reduced the need to own physical infrastructure and assets, allowing all manner of operators to construct and scale up digital platforms cheaply, quickly, and with relative ease (Carr 2008). On the other hand, achieving bigger scale and greater scope create barriers to entry: as the previous chapter spelled out, both direct and indirect network effects helped tech giants like Google and Meta amass vast datasets that they monetized much better than smaller rivals.

Property Rights Over Data

In the US, data is treated like oil, with ownership effectively vesting in those who "extract" it—companies that collect data claim rights to it. This approach is a pragmatic, Coasean-type solution to the immense transaction costs associated with personal data.[165] Consider that

[165] In Europe's "infrastructure-oriented" view, user rights are privileged under laws like the GDPR, and data is treated as a shared resource requiring stewardship. China, meanwhile, views

according to Coase (1960), in a world with low or zero transaction costs, the initial assignment of a property right is of little consequence for social welfare; the parties involved can simply bargain and trade the right until it is held by the party who values it most, ensuring a socially efficient outcome. Data should be no different: if individual users and digital platforms could negotiate and contract at zero cost, the data would inevitably gravitate toward the party capable of extracting the highest economic utility from it, regardless of whether the legal right was originally assigned to the user or the platform.

However, in both the real and digital world, where transaction costs are not negligible, there are significant barriers to such bargaining, effectively "locking in" the initial asset allocation. To maximize social welfare in this “second-best world,” a Coasean approach dictates that policymakers should assign the property right to whichever party values it the most and can generate the greatest amount of social welfare by possessing it. Therefore, in the non-zero transaction cost real and digital worlds, getting the initial assignment of rights “right” is imperative: assigning it “incorrectly” invites suboptimal outcomes—the resource will not be put to its most productive use and social welfare will be lower than it otherwise would be.

So now comes the million-dollar question: in the second best world we actually live in, who should the government assign the property right over data to, digital platforms or users?

---

data as a public good under state oversight, prioritizing national security and social governance over individual claims (see Chapter 12).

If property rights were assigned to each individual user for every bit of data they create, a viable market might never form, culminating in considerable deadweight losses.[166] Even if users were willing to sell their data, the transaction costs of identifying, metering, and tracking the creation and exchange of billions of fragmented data points would be insurmountable. Imagine the ungodly task of attributing revenue shares across countless contributors for a single algorithmic insight. Because a datum only becomes a "prediction capable resource" when refined alongside millions of others, any effort to assign separate "titles" to these slivers of data never effectively translate into a liquid market, trapping the resource in its raw state instead.[167]

Consequently, assigning the property right to the digital platform, the only entity capable of collecting, aggregating, and analyzing this resource at scale, is the Coasean path that unlocks the most social value.

---

[166] A deadweight loss is the loss in economic efficiency that occurs when mutually beneficial trades that would otherwise occur fail to take place.

[167] That does not discount that several creative ideas to make individual data ownership more tenable have been proposed by some data privacy advocates. The most prominent: the creation of “data unions” or “data cooperatives” where individual users would pool their data and the union would collectively bargain with tech platforms on their behalf for payment or better terms, solving the problem of individual data possessing “too little” value in isolation (see Posner and Weyl 2018). For this vision to reach fruition, however, it would likely require a new legal framework as well as companion innovations like “data fiduciaries”—entities with a legal duty to act in their clients' best interests—underpinned by blockchain ledgers and “smart contracts” that could both track data usage and distribute micropayments to users.

Several examples bear this prescription out. First, consider Google Maps: it is only after millions of location pings are cleaned, fused, and modeled that raw spatial coordinates metamorphize into a new asset with true commercial value—accurate, up-to-the-minute traffic predictions (Lau 2020). Second, take Meta. In the third quarter of 2024, the company's global revenue per user was roughly $12—and profit per user was just under $5 (Meta Platforms 2024). This figure, representing the value extracted from thousands of users' likes, posts, and interactions over three months, is less than half the price of a movie ticket. This illustrates that the value of any single data point—one click, or one "like"—is a microscopic fraction of a cent, making it economically impossible to price and sell individually. Third, consider Google Search: the billions of daily queries that users enter—many of which are unique or "long-tail" phrases—only generate value because Google can correlate them with billions of previous "click signals" to predict intent (Gomes 2017). On its own, a single user's search for an obscure medical symptom is a "sand-like" data point; it only becomes a powerful medical insight or a refined training input for an LLM when indexed against the collective search history of millions of others.

Nuts and Bolts of the US Approach

Because US copyright law does not protect raw facts (US Congress 1976b; see Feist 1991), and the country lacks a special intellectual property right for databases, a combination of legal agreements and technical measures remains the primary way platforms control and commercialize data. When a user signs up for a service, the "terms of service" act as a binding contract granting the company that collects the data broad rights over user generated data. That means that, practically speaking, "clickwrap" agreements—the "I Agree" boxes users must check

to access a service—force users to license away control to the companies collecting it (Fairfield 2017).

This practice is underpinned by the law and judicial system in several ways. The E-SIGN Act established that electronic contracts and signatures have the same legal validity as their paper-and-ink counterparts (US Congress 2000) and is reinforced by technical “access controls” that act as digital fences and by “trade secrecy” laws that protect a company's refined dataset as a proprietary corporate asset (Pooley 2015; US Congress 1996). While copyright does not protect facts, trade-secret law can protect curated datasets/compilations and query systems provided the owner can prove to a court that they employed reasonable technological and legal barriers to maintain secrecy (Compulife 2020).

Courts consistently held that if the “terms of service” are clearly presented by a digital platform or publisher, clicking the box is enough to bind a user to the rules, such as permitting the platform to collect and commercialize their personal information (see Facebook 2019). They also ruled that users don't always need to click "I Agree" to be bound by a contract; for example, if a company warns a data scraper about its terms and the scraper continues to access the site anyway, that continued use counts as acceptance of the rules (see Register.com 2004). These rulings not only enforced platforms' contractual control, but legitimized their de facto property

rights over collected data, allowing them to both treat aggregated datasets as proprietary assets for commercialization and exclude unauthorized access to their sites.[168]

I hasten to emphasize, however, that digital platforms' rights over user data are bounded in several important ways, including the FTC's authority to police "unfair or deceptive" practices, which prevents companies from using overtly misleading or hidden terms; narrow, sector-specific laws for sensitive information, such as HIPAA for health data; and broader state-level laws like the California Consumer Privacy Act (CCPA/CPRA).[169]

Furthermore, the practical reach of these regulations extends far beyond their jurisdictional borders. Because California represents such a massive share of the US market, digital platforms often find it technically and legally inefficient to maintain separate data infrastructures for Californian and non-Californian users (see Vogel 1995; Chander et al. 2021). Consequently, these platforms often extend CCPA-compliant rights—such as the "Do Not Sell My Info" button—to all American users as a global default (Pardau 2018). Similarly, the

---

[168] Platforms have historically relied on anti-hacking statutes—like the Computer Fraud and Abuse Act (CFAA)—and digital trespassing doctrines to block automated scraping, particularly when a bot bypasses a login screen or other security measure (see, e.g., eBay 2000). However, courts significantly narrowed CFAA liability for scraping publicly accessible data in 2022 (see hiQ Labs 2022).

[169] On the FTC's role, see Solove and Hartzog (2014). On the emergence of state-level regimes like the CCPA, see Rothstein and Tovino (2019).

"Brussels Effect" (Bradford 2020) means that the European Union's stringent General Data Protection Regulation (GDPR) standards often trickle down into the global operations of US tech giants: To avoid the legal peril and engineering complexity of ring-fencing data by region, platforms frequently upgrade their baseline privacy terms to meet European standards, thereby granting American users a higher level of de facto protection than federal law alone would require (ibid).

Creating a Market for Data Creation and Exchange

Given that data's profitability hinges on massive scale, a market eventually emerged trading in aggregated "personal insights" rather than individual data points. This marketplace functions as a high-frequency "ad-tech" stack where billions of data points are bid upon in milliseconds. At its core, the market consists of data brokers who ingest raw signals, "Data Management Platforms" that categorize these signals into "audience segments," and automated exchanges where these segments are auctioned off to advertisers via "Real-Time Bidding." Rather than a static exchange of files, this market is a dynamic, real-time plumbing system that matches a specific user's profile to a commercial intent at the exact moment a webpage loads (UK CaMA 2020; Hwang 2020).

The emergence of this industrial-scale exchange was not inevitable; besides the legal and regulatory landscape I outlined above, consisting of a "contract-over-property" framework that favors data extractors, the absence of a federal database right, and the enforcement of "clickwrap" licensing agreements, it was made possible by a convergence of three innovations. First, the democratization of "cloud computing" allowed even small startups to rent vast, on-demand storage and processing power without the upfront cost of building data centers (Mills

2021). Second, advances in “big data” analytics and “machine learning” leveraged this infrastructure to sift through immense volumes of information, generating the predictive behavioral insights advertisers desired (McAfee and Brynjolfsson 2012). Finally, new privacy-preserving protocols—such as “differential privacy,” which adds statistical noise to mask identities (Dwork and Roth 2014), and “homomorphic encryption,” which allows data to be processed without decryption (Gentry 2009)—provided a secure mechanism for these brokers to trade sensitive profiles without exposing raw user information.

The Early, Permissive Era

For many years, the digital data landscape was highly fragmented but defined by a central paradox: it was simultaneously a collection of fiercely protected "walled gardens" and a wide-open plain for data tracking. On one hand, each tech giant built its empire on a core, proprietary dataset that was kept inaccessible to its rivals; Apple’s iMessage history, Amazon’s purchase data, and Google’s search queries remained siloed within their respective fortresses. However, the walls of these gardens were made porous by the permissive tracking technologies of the era.

While during the 2000s and early 2010s most individual vendors held only isolated pieces of a user's data, third-party cookies and tracking pixels embedded across millions of websites meant that a user's physical movements and digital activities were transformed into a single digital identity that could follow them across devices and platforms, enabling advertisers to engage in "hyper-targeting" based on their real-world behaviors and predicted future needs

(see FTC 2017b).[170] Giants like Google and Meta exploited their vast ecosystems and ubiquitous ad networks—systems that aggregate digital advertising inventory from thousands of publisher websites to sell to advertisers centrally—to assemble comprehensive, monetizable user profiles. Similarly, while a vast ecosystem of mobile apps, such as the popular WeatherBug, often collected granular location data as a condition of use (Valentino-DeVries et al. 2018), internet retailers could passively track customers' movements between virtual "stores" because their smartphone's Wi-Fi would constantly broadcast unique hardware identifier, or "MAC addresses" (Clifford and Hardy 2013).

Consider the example of a fictional user named Sara during the mid-2010s.[171] On her Android phone, she frequently used the WeatherBug app, granting it location access. This app (owned by GroundTruth) would log her location as she commuted. Sarah also spent hours browsing for new recipes on food blogs, frequently clicking on affiliate links to kitchen gadgets, and "liking" posts on Facebook about healthy eating. She also occasionally upvoted restaurant reviews on Yelp and would check her Gmail compulsively throughout the day across devices.

Each of these activities—her physical location, web browsing, social media engagement (liking, posting), and Gmail and app usage—generated data. GroundTruth gathered Sara's anonymized location data from WeatherBug. Google, through its advertising network

[170] "Cookies" are small files with a unique identifier that follow a user across websites, while "tracking pixels"—invisible $1x1$ images—allow platforms to record specific actions (like a purchase or a click) and tie them back to the user's centralized profile.

[171] For a similar illustrative approach from an industry perspective, see Apple 2021a.

(DoubleClick, AdX) and Chrome browser, collected data on her food blog browsing and clicks. Facebook collected her "likes" and posts on its platform.

This data was then leveraged for advertising. When Sara visited a website with available ad slots (e.g., a news site), GroundTruth might sell ad placements to a kitchen appliance manufacturer, targeting her based on location data after inferring a strong interest in cooking. Google's ad exchange would run an auction, allowing advertisers to bid for slots on those food blogs, with a kitchenware company potentially buying the slot, targeting Sara based on her browsing history. Facebook, meanwhile, would digest Sara's history of likes and serve cooking-related ads directly into her news feed.

Sara's data stream was also used by app developers—during this time, Google allowed them to gain access to SMS messaging data, call log data, some forms of contact data on Android phones, and Gmail messages.[172] Therefore, a "smart" caller ID app optimized to stop spam calls in their tracks could request access to her call logs and entire contact list. It would then upload this information to its own servers, combining it with data from millions of other users to create a massive, crowdsourced phone directory. The app could then identify incoming calls from unknown numbers by matching them against this global directory.

---

[172] While Google's developer agreement prohibited exposing a user's private data to anyone else "without explicit opt-in consent from that user" and its rules also barred developers from making permanent copies of user data and storing them in a database, before 2019 developers used to find ways to get Gmail users to give up access to their inboxes without clearly stating what data they collected (see MacMillan 2018).

In short, the permissive era was defined by a fundamental decoupling of data creation from data control. Because the "transaction costs" of tracking were lowered to near zero by a combination of weak technical fences and a "contract-over-property" legal framework, the property rights over Sara's digital exhaust effectively vested not with her, but with the industrial-scale "refiners" capable of aggregating her fragmented activities into a coherent, monetizable asset.

The "Privacy" Era Emerges

The "Wild West" data collection landscape that defined the internet over the first decade and a half of the 21st century was progressively dismantled—not by a single law, but by a series of deliberate privacy-enhancing changes made by Apple and Google.[173] While the third-party cookie was the foundational technology for cross-site profiling for over two decades, the industry began to move away from it around 2024, presaged by the popularity of web browsers, such as Safari and Firefox, which had long blocked such trackers by default (Wilander 2017; Mozilla 2019). Even Google Chrome, the market leader in this space, transitioned toward a model of "user choice," prompting individuals to decide whether to allow tracking rather than permitting it by default (Chavez 2024).

[173] To be sure, this model was significantly altered in the crucible of a new era of privacy regulation brought about by sweeping laws like the GDPR and CCPA. These laws legally mandated data minimization and granted users the right to delete their information, transforming privacy from a matter of "internal diligence" into a high stakes engineering requirement for global platforms (Bradford 2020; Chander et al. 2021).

Perhaps the most important changes occurred at the operating system level. To combat passive Wi-Fi tracking, both iOS and Android implemented "MAC address randomization," which broadcasts a temporary, fake address when scanning for networks, making it impossible to track a specific device (Martin et al. 2017). Both also introduced "granular location permissions," forcing apps to ask users to choose whether they can access location "Always," "Only While Using the App," or "Never," and began sending periodic reminders about background tracking (Federighi 2019; D'Silva 2020).

Indeed, while the introduction of Apple's "App Tracking Transparency" framework in 2021 bolstered the company's sterling reputation for data privacy by requiring apps to get explicit, opt-in consent from users before tracking them across other companies' apps and websites (Apple 2021b), Google had already undertaken a series of similar changes. By that time, it had already restricted third-party access to sensitive data, instituted a near-total ban on the human reading of user data, and mandated rigorous security reviews for apps (Smith 2018).[174] Google also allowed users to opt out of personalized ads; furthermore, it discontinued the practice of scanning Gmail users' messages to serve ads to them in 2017 (Bergen and Surane 2017)—and developed the so-called Privacy Sandbox, a set of technologies designed to allow for

[174] Following the 2018 announcement of "Project Strobe," Google implemented a strict new verification process in 2019 that severely limited which apps could request access to Gmail and explicitly prohibited the practice of human employees reading user emails except in rare, consented-to cases (Smith 2018).

interest-based advertising without tracking individual users across different websites (Chavez 2024).[175]

Social networks also evolved in line with these privacy changes. Facebook, for example, adapted by shifting its focus away from broad data sharing and toward two primary sources of information. The first was its own “on-platform data,” the rich web of connections and interests known as the Social Graph. The second was “advertiser-provided first-party data,” where businesses could use tools like the “Custom Audiences” feature to target their existing customer lists.[176] This occurred after it abandoned its so-called “Partner Categories Program,” which

---

[175] For example, its Chrome browser categorizes a user's interests into broad topics like "Fitness" or "Travel" and shares only those general topics with advertisers, not their specific browsing history. Moreover, when a user opts out of personalization, their data is no longer used to build an advertising profile or show them targeted ads. However, at the time of this writing, Google continues to collect its users’ search and browsing history (unless separately disabled) for other purposes, such as improving its core search algorithms, enhancing security, and for non-personalized, "contextual ads" like showing an ad for a car on an automotive review website (Google 2025a).

[176] Custom Audiences is an advertising feature that allows a business to use its own first-party data to find and target its existing customers on Meta's platforms. Historically, this was often done by tracking website visitors with the Meta Pixel, a tool that relied on third-party cookies; businesses later took to uploading a list of its customer information, such as email addresses or phone numbers, which Meta then matched to its user profiles in a "privacy-safe" way, creating a "custom audience" for the advertiser to target (Facebook 2025).

allowed third-party data brokers to directly integrate their vast stores of offline data—such as purchase histories and financial information—with Meta's on-platform data (Constine 2018). Facebook also shut down its broad developer access in the wake of the Cambridge Analytica scandal, which erupted after a third-party quiz app harvested the personal data of millions of users in 2014 to create political profiles (Cadwalladr and Graham-Harrison 2018).[177]

By 2025, as users gained more direct control over who could see their location and how that data is used, digital platforms inhabited a fundamentally altered landscape. While location-based advertising still existed, the era of unfettered, opaque, cross-app tracking by third-party data brokers had been severely curtailed. Ad targeting came to depend on "work arounds" like directly uploaded customer lists (Darveau-Garneau et al. 2022).

Markets for AI Training Datasets Thrive

While the advertising industry was fixated on tracking user behavior, a different kind of asset was quietly accumulating in the background: the content of the web itself. Billions of articles, essays, and conversation threads were scraped and organized into massive, publicly available training repositories, the most famous of which is the Common Crawl corpus. For early AI labs, these vast, pre-crawled libraries were an invaluable public resource, providing a ready-made source of diverse, real-world language at almost zero marginal cost that allowed them to train foundational models without having to build their own costly web-scraping infrastructure (see Dodge et al. 2021).

---

[177] While Cambridge Analytica was first hired by Ted Cruz's 2016 presidential campaign, these profiles were most famously exploited by Donald Trump's successful campaign that year to serve hyper-targeted ads.

Moreover, the decline of third-party cookies and Apple's App Tracking Transparency, which limited the effectiveness of pixel-based tracking, did not slow down the race to train AI. The privacy era still produced just as much rich and humanlike data as before, if not more of it (IDC 2023); the new rules simply changed where that data was generated. Instead of relying on third-party cookies to track users across the open web, platforms focused on maximizing engagement within their own "walled gardens" (see LiveRamp 2024). The explosion of on-platform, user-generated content—from short-form videos on TikTok and Instagram Reels to integrated e-commerce via “Shops”—created an even deeper and more concentrated well of first-party data that was immune to cookie deprecation (Srinivasan 2019).[178]

To understand the origin of the vast “data reserves” refined into the high-octane fuel that powered the AI revolution, I now turn to the legal and regulatory framework that made their extraction possible.

SECTION 230'S LIABILITY SHIELD AND RISE OF PERSONALIZED PLATFORMS

Section 230 of the 1996 Communications Decency Act—famously encapsulated in the "26 words that created the internet" (Kosseff 2019)—immunizes online intermediaries both for

[178] After fully exploiting the public web, AI developers began seeking higher quality, specialized, and proprietary data to give their models a competitive edge. Traditional content creators capitalized on this demand to generate a new revenue stream: news publishers, stock media libraries, and platforms such as Reddit entered licensing partnerships with AI labs for access to their content and data (Brougher 2025) Chapter 9, I discuss a new industry of “synthetic data providers” that offer artificially generated, perfectly labeled datasets for training AI in niche areas where real-world data is scarce or sensitive.

hosting third-party content and for voluntarily removing content they find objectionable (US Congress 1996). This protection for moderation decisions meant that even aggressive algorithmic shaping of what each user sees—promoting some posts, burying others—would not typically expose platforms to liability from speakers claiming censorship or viewers claiming the platform negligently suppressed or de-prioritized content (Kosseff 2019).

Ensuing legal cases then transformed Section 230 into a sweeping liability shield, empowering platforms and AI labs to (i) freely experiment with personalized, algorithmically driven recommendation systems that personalize users' experience to juice engagement, representing AI 1.0; (ii) foster the growth of massive, scrapeable human-interaction datasets; and (iii) train LLMs using the billions of words and images caught in this legal net, representing AI 2.0.

Early court decisions cementing broad immunity

Kosseff (2019) canonically demonstrates that judicial interpretations of Section 230 routinely protected platforms against lawsuits, transforming it into an immunity forcefield against torts.[179] While platforms might bear responsibility for how their design choices affect users (especially

[179] While Jeff Kosseff's definitive history, "The Twenty-Six Words That Created the Internet" (2019), chronicles the statute's first two decades, this section both reviews his main insights and goes well-beyond them, focusing in particular on the "Algorithmic Era" of case law that emerged after that book's publication. Specifically, further ahead I examine post-2019 landmark decisions like Lemmon v. Snap (Lemmon 2021), Domen v. Vimeo (Domen 2021), and Gonzalez v. Google (Gonzalez 2023) that have tested the limits of legal immunity regarding product design, algorithmic amplification, and the rise of AI-driven recommendations.

minors), that is not the case for what users say. As of the mid-2020s, companies enjoyed broad immunity for users who harmed others while on their platforms, while carving out narrow slices of unprotected conduct when the platform's fingerprints had clearly left their own signature.

In the landmark Zeran (1997), the court created a significant precedent by deciding that AOL, one of the first internet platforms, was not liable for failing to remove defamatory posts about the plaintiff, even after being notified of their falsity. So long as a third party (user) created the "harmful" content, rather than the platform, which merely hosted it, then it was the former who was responsible for it. It did not matter if the allegation was negligence, defamation, or any other complaint about that content.

Courts across jurisdictions subsequently embraced this broad approach. In practical terms, Section 230 meant, for example, that while a business alleging defamation could sue a post's author, it could not go after the message board or app where it appeared. Therefore, courts summarily tossed out similar lawsuits spanning claims from defamation to discrimination (see Kosseff 2019).[180]

---

[180] Platforms cannot simply recast every lawsuit as an attack on third-party content, however. If the plaintiff can persuade a court that the injury was caused by the platform's own features or conduct (and not by the speech of another user), or that it failed to uphold a duty to its users, Section 230 may not apply. Consider Lemmon (2021), a case where parents sued after their teens died in a crash while using Snapchat's "Speed Filter" feature. Though Snap invoked Section 230, arguing the case concerned user-generated snaps, a court allowed the suit to proceed, holding that the claim centered on Snap's product design—creating a feature that was akin to defectively designed tool by allegedly incentivizing dangerous speeding—not on publishing user content.

For example, in Doe (2016), victims of sex trafficking alleged that Backpage (an adult ads platform) designed its site to actively facilitate this illegal activity. While the court empathized with the victims' plight, it upheld Section 230 immunity, bluntly observing that the law compelled this outcome "even if enforcement of the immunity results in many horrible things going unaddressed" (ibid: 29).[181] The court reasoned that Backpage's choices in structuring the classified ads that appeared on its site (what categories to allow, how to screen posts, etc.) were covered by Section 230, since Backpage had no material part in crafting the

---

The main takeaway: because the duty not to design unreasonably dangerous products exists *independently of harmful user speech or the platform's content moderation functions*, it falls outside Section 230's immunity shield.

[181] While Section 230's protection is broad, courts have carved out exceptions when platforms cross the line from neutral hosts to content creators or make specific promises. In Fair Housing Council (2008), a court denied immunity to Roommates.com, a platform that matches housemates, because the site required users to disclose protected characteristics for housing searches, thereby "materially contributing" to illegal discrimination rather than passively hosting user content. Similarly, in Barnes (2009), a court allowed a "promissory estoppel claim"—a legal doctrine allowing someone to enforce a promise even without a formal contract when they reasonably relied on that promise to their detriment—to proceed when Yahoo failed to fulfill an employee's promise to remove revenge porn, establishing that platforms can face liability for breaching their own commitments to users.

actual content of the ads—thus shielding the platform from civil liability for the illicit user posts.[182]

Likewise, attempts to hold online matchmaking services and marketplaces liable for criminal transactions (guns, drugs, etc.) and behaviors enabled by site design have mostly failed on similar grounds: courts conclude that the alleged design flaws (like insufficient user screening or anonymous listings) are too entwined with decisions about hosting third-party listings, which 230 protects. For example, when a victim of severe harassment sued the dating app Grindr for negligently designing its app without safeguards to prevent impersonation and abuse, the court found that, despite the "product design" labels, the harms flowed from the content of a fake user profile—squarely third-party content and therefore protected by Section 230 (see Herrick 2018).

Not only did Section 230 immunize platforms against doing nothing (leaving content up), but it also protected them for taking it down. By explicitly protecting "Good Samaritan" filtering, Section 230 furnished them with wide latitude to remove or curate content as they saw fit (Kosseff 2019).

Domen (2021) underscores this second-order freedom, even when the speech in question is protected against government censorship. There, Vimeo, a video platform, removed user-uploaded videos that promoted "sexual orientation change efforts" (a.k.a. conversion therapy), citing its content policies. The plaintiff sued, alleging religious discrimination and unfair

[182] Congress responded to the Backpage controversy by enacting FOSTA-SESTA in 2018, which carved out an exception to Section 230 for content that promotes or facilitates sex trafficking—a rare legislative narrowing of platform immunity (US Congress 2018).

censorship. The court, siding with Vimeo, flatly held that Section 230 immunized the platform's decision to ban the videos, finding that it considered the videos objectionable under its policies.

The upshot is that platforms could go so far as to remove content outright—banning users or deleting posts that violate standards for reasons like user experience or brand safety. But Section 230 caselaw also allowed for everything in between too. For example, platforms could experiment with different moderation strategies (from totally open forums to heavily curated communities) without incurring unique legal risk, so long as the content was user-supplied.

Algorithmic amplification as "publisher" activity.

A paramount question emerged during the AI age: whether the Section 230 immunity shield that digital platforms enjoy regarding third-party content extends to the algorithmic curation and personalized recommendation of that content. On the one hand, as platforms moved beyond static message boards to highly personalized feeds powered by algorithms that enjoyed trade secret IP protections, plaintiffs argued that choosing which content to push to a particular user (and in what order) is a form of active editorial involvement that should void 230 immunity—especially if the platform's algorithms tend to amplify harmful material. On the other hand, courts' instincts were traditionally to treat recommendation algorithms as just another "neutral tool" helping disseminate third-party content—with some few important exceptions.

Force (2019) embodies the now dominant approach. In that case, victims of terror attacks alleged Facebook's News Feed and friend suggestion algorithms helped radicalize users by promoting Islamic State (ISIS) propaganda and connecting extremists to each other. The court rejected this argument and upheld immunity, holding that Facebook's algorithmic systems for organizing and recommending content were simply extensions of its role as a platform hosting other parties' content. It emphasized that every platform must make choices about how to display

and prioritize user posts—doing so doesn't automatically transform Facebook into the developer of the underlying content. In other words, "ranking, sorting, or matching content" is fundamentally a distribution function covered by Section 230.

Two subsequent cases extended this reasoning to algorithmic curation.

In Dyroff (2019), the mother of a young man who died from a fentanyl overdose alleged that a social networking platform, Ultimate Software, was liable because its automated friend and group suggestion algorithms connected her son with a drug dealer. The court, however, rejected this claim and affirmed Section 230 immunity, reasoning that the recommendation algorithms were "neutral tools" for organizing user-provided data that served the function of connecting different users and did not materially contribute to the illegality of the drug seller's posts.

Similarly, when families of terrorism victims sued YouTube (Google) in Gonzalez (2023), claiming its targeted video recommendations aided ISIS recruitment and thus violated the Anti-Terrorism Act, the Supreme Court declined to address the "230 question." Instead, it held that the plaintiffs failed to state a valid claim under the Anti-Terrorism Act regardless of any immunity issue—leaving algorithmic amplification concerns unresolved at the highest judicial level.

However, the "neutral tool" defense began to show significant fractures leading up to 2025. Outlier rulings in cases like Malwarebytes (2020) and Anderson (2024) signaled a new judicial willingness to pierce the Section 230 veil when platforms acted in "bad faith" against competitors or when their algorithms aggressively promoted deadly challenges to minors. These decisions laid the critical legal groundwork for the heightened liability scrutiny that would later greet the arrival of Generative AI. Importantly though: these fissures emerged only after the

digital economy had already achieved escape-velocity, rendering this somewhat of a moot point. For the formative decades of the Fourth Industrial Revolution, the 230 shield did its job as intended.

AI Enhanced Moderation Also Flourished

Under the aegis of Section 230, platforms also down-rank or filter "low quality" or rule-breaking posts algorithmically by, for example, deprioritizing misinformation or hate speech in feeds —a practice critics often deride as "shadowbanning" (Nicholas 2022). Years of iterative tuning of content policies and algorithms meant that early laissez-faire approaches gave way to stricter rules against harassment, hate speech, or disinformation, all enforced by a mix of human moderators and AI filters, especially as public and political pressures mounted (Klonick 2018; Gillespie 2018). Section 230, empowered by AI, meant each platform could calibrate these choices to its audience and business, rather than a one-size legal standard (Klonick 2018).

Why This Matters for the Fourth Industrial Revolution

By unleashing an era of platforms built on user-generated content, Section 230 became the legal bedrock for countless websites, public forums, and social networks (Kosseff 2019). Shorn of the fear of being held liable for any harm caused by their users, these platforms were empowered to host a vast spectrum of third-party content on a massive scale. That includes users' posts, comments, reviews, photos, videos, wiki articles, and even code repositories.

Protected by Section 230, platforms could also experiment with algorithmically generated recommendations based on user patterns and feedback to provide personalized, engagement-optimized services. Concretely, rather than accumulate armies of attorneys and pre-publication reviewers to heavily vet or suppress user content to ward off lawsuits, tech companies could instead invest in growing their platforms by rolling out community-driven features quickly.

Several examples come to mind. Facebook, for instance, introduced algorithmic News Feed ranking around 2009 and continued refining it thereafter, confident that tweaking the order of user posts wouldn't open it up to defamation or negligence suits over which user content got amplified (Facebook 2013). Likewise, YouTube's famous recommendation engine (responsible for over 70 percent of the platform's viewership at the time of this writing) suggested videos through its "Up Next" autoplay suggestions—even controversial or extreme ones—without Google fearing that each automated suggestion would be treated as its own editorial act (Solsman 2018). The same was true for TikTok's algorithmically driven "For You Page" (Zulli and Zulli 2022).

Thanks to an underlying legal assumption that platforms were mere intermediaries and should not be treated as speakers of user content, by the early 2010s the largest digital platforms delivered user generated content and interactions in the billions each day while collecting detailed behavioral data on what each person clicked, liked, shared, or ignored. Algorithms optimized engagement on vast stores of user content—and in doing so, generated unprecedented volumes of data about human communication, preferences, and behavior, flowing from Facebook's daily active users, YouTube's hours of video uploaded per minute, and Reddit's endless comment threads. In essence, Section 230 allowed what Tim Wu (2016) derisively termed “the attention economy” to flourish and social media networks to become global behemoths—because they could scale user engagement without proportional increases in legal exposure or moderation costs (Gillespie 2018).

THE OPEN WEB AND DATA FOR TRAINING AI MODELS

Most important for our purposes, vast repositories of human expression came to be freely available for the taking after users posted them on the web. AI model trainers pounced,

exploiting the publicly available Common Crawl, an evolving snapshot of the digital commons composed of trillions of words from news articles, forum discussions, product reviews, blog posts, and open-source code repositories. Abundant, diverse, and conversational data was ripe for the taking.

This archive was itself a byproduct of the positive-sum relationship forged between content creators and online platforms between the mid-1990s and early 2020s. Under a widely accepted system known as robots.txt—a small file on a website that tells web crawlers which parts of the site they are allowed or not allowed to access—publishers openly shared their entire web collections, including news archives, blog posts, forum discussions, and technical documents (Koster 1994; IETF 2022). Search-engine crawlers like Googlebot and Bingbot, automated programs that browse the internet to index content, hoovered up this content. In return for this open access, search engines acted as a powerful distribution channel, sending billions of visitors back to these publishers (Bell and Owen 2017). This steady stream of referral traffic generated valuable advertising impressions and drove new subscriptions, creating vital revenue for creators and bankrolling publishers' continued investment in original content (Nielsen and Ganter 2022). However, this cooperative equilibrium rested on the assumption that crawlers were bringing visitors back to the source, rather than harvesting content to train a technological replacement for the publisher themselves.

Technologies, Norms, and Laws that Supported the Open Web System

The ability of AI modelers to freely harvest the web was not the result of a single law, but of two distinct institutional eras that functioned on the same logic: an economic incentive supported by a permissive legal framework.

In the first era, the Search Era, content creators and platforms engaged in a "positive sum" relationship defined by a cooperative bargain. The mechanics of this era relied on robots.txt as a technological handshake (Koster 1994; IETF 2022), signaling which parts of a site were accessible to crawlers. The economic incentive was a clear win-win: in exchange for access, search engines functioned as a distribution channel, sending billions of visitors back to publishers (Bell and Owen 2017). This referral traffic bankrolled the creation of original content through advertising and subscriptions, creating a legal reality rooted in tacit consent (Nielsen and Ganter 2022). As long as the traffic flowed back to the creators, the legal question of "taking" data was irrelevant because the economic benefit was self-evident.

As the AI revolution began, the system transitioned into the AI Era, and the mechanics shifted from distribution to extraction. AI modelers "pounced" on the digital commons, using archives like the Common Crawl to harvest trillions of words without any intent of sending traffic back to the source. The economic incentive shifted entirely to the modelers; by treating the web as "statistical fodder," they could train LLMs at near-zero marginal cost, effectively bypassing the need for bespoke licensing agreements.

When the cooperative handshake of the previous era failed, developers relied on the legal weakness of the system to maintain their momentum. This new legal reality utilized the fact that robots.txt carries no statutory force and leaned on the hiQ Labs (2022) precedent, which affirms that scraping public data does not violate anti-hacking laws (see US Congress 1986b). Finally, the Bartz (2025) and Kadrey (2023) rulings, which I detail below, provided the "copyright blessing," declaring that using this data to learn patterns—rather than to copy text—was a protected "transformative" use.

Key Legal Rulings About AI and Copyright

In Bartz et al. (2025), a court held that training AI models on lawfully acquired copyrighted books constituted fair use. Anthropic's use was deemed "exceedingly transformative" because the books were employed only to develop an AI's predictive capabilities, not to reproduce their narrative or expressive content: Where a traditional reader consumes a book and then re-expresses its ideas, Anthropic instead used the texts as unlabeled statistical fodder to optimize complex probability models, a use that was far removed from the expressive or commercial markets the authors occupy.[183] In other words, the court found it would be "unthinkable" to require payment each time a human reader —or, by extension, a machine— recalls or builds upon a text in speech or writing.

In Kadrey et al. (2023), the court granted Meta's motion to dismiss, finding that the plaintiffs, a gaggle of famous authors of popular books, failed to show substantial reproduction or market harm from its use of their works when the company trained its Llama models. The judge argued that although Meta's LLMs may echo snippets of their training corpus, and therefore snippets of published works, they were unable to reliably regurgitate large, recognizable blocks of text from any one plaintiff's work; consequently, incidental recitations were too trivial to support a cognizable claim or threaten the authors' markets. Second, while the plaintiffs argued they had a right to be paid for use of their books as AI training feedstock, the

[183] However, the court drew a bright line at Anthropic's admitted "pirating" of more than seven million titles via unauthorized repositories such as Library Genesis that distributed copyrighted works without authorization. The judge ruled that while the act of training on copyrighted material may be protected, illegally obtaining that material violated copyright law, even if the subsequent training might have been permissible had the books been obtained legally.

judge concluded that no such licensing market currently existed, and if it did, the complaint offered no evidence of its scope or their lost opportunities.

How AI Model Development Blossomed in this Ecosystem

Empowered by the technology, norms, and laws outlined above, early AI labs bypassed the need for bespoke licensing agreements or custom crawlers. Instead, they tapped into massive, pre-crawled datasets that provided trillions of tokens at almost zero marginal cost.[184] OpenAI, lacking its own proprietary user-data engine like Google's, built its ChatGPT models by tapping into the vast "digital commons" of public web data. Common Crawl was the largest component of GPT-3's training mix, at 60%; it provided petabyte-scale monthly snapshots that collected hundreds of billions of words from academic papers, enthusiast blogs, developer forums, product manuals, reviews, and news stories (Brown et al. 2020).[185]

OpenAI also drew on other sources to train this model. This included two large book collections (totaling 67 billion tokens) and the entire English Wikipedia (see Brown et al. 2020;

---

[184] A “token” is the basic unit of text that a large language model reads and processes, which can be a full word like "apple," a part of a word like "un-" or "-ness," or even a punctuation mark. Generally, for English text, one token is roughly equivalent to four characters or about 75% of a word, and this process of converting human language into a sequence of numbers that the model can analyze and use to make predictions is called “tokenization.”

[185] A petabyte is 1,000 terabytes or a million gigabytes. Recrawl schedules, determining how often websites were rescanned, varied from hourly for rapidly changing forums to nightly for major news outlets.

see also Kang et al. 2024).).[186] However, to ensure GPT-3 learned from high-quality, relevant, and current content, the lab also curated datasets based on signals of human preference. Its influential WebText2 dataset, for example, was built by scraping only those web pages that were linked in posts on the social media site Reddit and had received a high number of "up-votes"(Radford et al. 2019).[187] This practice of using social signals to surface valuable content meant that the mobile-first, user-generated data from GPS-enabled platforms became a crucial ingredient. Content shared widely on social platforms—such as local news blogs, travel diaries, and review sites—supplied the unique regional language, specialized jargon, and “long-tail” knowledge that gave the model its breadth. While exact figures are difficult to verify, some analyses suggest a significant portion of the text in these large web datasets comes from mobile-sourced user content, which retains implicit geo-signals like mentions of local landmarks precisely because explicit location data is stripped out (see Radford et al. 2019).[188]

---

[186] In terms of the data consumed, this meant 3 billion tokens from English-language Wikipedia pages and 55 billion tokens from two datasets believed to contain text from millions of published books (Brown et al. 2020; see also Kang et al 2024 and Gao et al. 2020 regarding "The Pile").

[187] In terms of the data consumed, this meant 19 billion tokens from web pages that were linked from Reddit posts with 3+ upvotes (see Brown et al. 2020; see also Radford et al. 2019 for the foundational WebText methodology).

[188] According to Radford et al. (2019), explicit geo-coding is minimal: per dataset audits, it is likely less than 5% of tokens. This lack of explicit metadata highlights how models must rely on implicit linguistic signals—such as mentions of local landmarks or regional dialects—to acquire geospatial reasoning.

This mobile-driven influx not only ballooned datasets to trillions of tokens but also enhanced temporal and spatial diversity and helped LLMs implicitly learn geospatial reasoning and real time adaptability. Without the need for EXIF tags, models were able to acquire knowledge about temporal, spatial, and sensory details.[189] This phenomenon also exposed models to real-time, context-rich narratives in plain, conversational English that captured evolving cultural nuances and on-the-go interactions. Think, for example, of Instagram captions describing "the best tacos in Austin" or a Yelp review of a "crowded café in downtown Tokyo during rush hour".

Google's Android operating system brought this app-driven data generation to a truly global scale. Its open-source model and dominance in the global smartphone market catalyzed an impressive volume of geo-tagged and dynamic data. While this was all part and parcel of Google's enduring effort to collect vast amounts of interactive data to refine its search and advertising capabilities, it also yielded a cache of AI training data of unparalleled scale and richness. I now turn my sights to explaining how this all came about.

THE GOOGLE FLYWHEEL

[189] EXIF stands for Exchangeable Image File format, a standard that automatically embeds metadata, or descriptive tags, directly into an image file when it is captured by a camera or smartphone. This metadata typically includes precise technical details such as the exact GPS coordinates and timestamp of where and when the photo was taken, as well as information about the camera itself, such as the make, model, and the specific settings used (e.g., aperture, shutter speed, ISO).

Google's epic flywheel did not emerge fully formed like Athena from Zeus's brow.[190] In an explicit effort to better monetize the world's largest stream of "intent" data, the company spent over two decades maniacally assembling a single, integrated stack. Said stack allowed Google to plan, target, run, and measure ad campaigns across multiple formats and devices (see Argenton and Prüfer 2012; Ramaswamy and Muret 2018; CMA 2020). The company's comprehensive indexing acted as a powerful incentive for publishers, from news organizations to individual bloggers, to create and digitize more content, knowing that visibility in search results was the primary gateway to a global audience. The ensuing explosion of online content, in turn, generated billions of daily page views—valuable digital real estate where ads could be placed.

To harness this dynamo, Google spent years developing and interconnecting a suite of indispensable, yet free, tools designed to embed its monetization and analytics engine into the fabric of the open web (UK CaMA 2020). Within this ecosystem, "Google AdSense" made it simple for any website owner to make money off their traffic by placing targeted ads. "Google Analytics" became the standard for measuring that traffic, which in turn gave Google an anonymized, bird's-eye view of user behavior across the entire web. Moreover, to consolidate its

[190] This paragraph synthesizes the CaMA's account of advertising-funded platform ecosystems and open-web advertising intermediation (CaMA 2020), Google's description of AdSense as publisher monetization through ad matching and auctions (Google n.d.), evidence on the widespread adoption of Google Analytics (W3Techs n.d.), and scholarship documenting Google's publisher tools as mechanisms of broad data collection beyond Google's first-party services (Schmidt 2018).

ad technology brands, Google relaunched "Google Ad Manager" (GAM)—which combined DoubleClick for Publishers and AdX—and the "Google Marketing Platform".

As I outlined in the previous chapter, this was all made possible by the company's strategic acquisitions. Google's 2007 purchase of DoubleClick gave it control over the desktop display advertising market, and just two years later, it acquired AdMob to dominate the rapidly growing mobile advertising space inside apps. In parallel, Google acquired the two platforms that would become the largest funnels for its data flywheel: the 2005 acquisition of Android put Google's services at the core of the global mobile ecosystem, while the 2006 purchase of YouTube secured its dominance over online video and provided a vital source of multi-modal data. Finally, its 2014 acquisition of the AI research lab DeepMind was about securing world-class talent and IP, granting Google a foundational advantage in the race to develop AI.

To ensure this flywheel spun on every device, Google systematically extended its reach beyond its own website through three pillars: its mobile operating system, its web browser, and a strategic partnership with its biggest rival. The company's Android operating system placed Google Search and other company services such as Gmail, Google Maps and YouTube as the default on the majority of the world's smartphones (Campbell-Kelly et al. 2015).[191] Moreover,

[191] Google used to scan Gmail users' messages to serve ads to them—a controversial practice that was part of the service from its 2004 launch until it was discontinued in 2017 (Bergen and Surane 2017). Henceforth, Gmail users see ads above their emails that are personalized based on other data Google collects, including their location, internet searches, and web history (see Schmidt 2018).

Google's Chrome browser.[192] Finally, Google paid Apple billions of dollars annually to be the default search engine on the iPhone, ensuring it captured the valuable traffic from its chief competitor's user base (Nylen 2023). This deal between unlikely bedfellows effectively blocked Microsoft's Bing from gaining the necessary scale on mobile to become a viable competitor to Google (see UK CaMA 2020).

These interlocking strategies produced unprecedented dominance across multiple digital markets. By 2024, Google consistently accounted for roughly 90 percent of online search queries in the United States, giving it extraordinary influence over the flow of information on the internet (StatCounter 2024). Chrome became the dominant gateway to the internet, capturing over 70 percent of the global browser market share by 2025 and funneling an immense volume of search queries and browsing data directly into its ecosystem (StatCounter 2025). Google's reach also encompassed online video: YouTube was the leading US ad-supported digital video platform by user engagement and viewing time, serving as a central hub for creators and advertisers alike (Statista 2024a). Google's dominance extended to mobile operating systems as well, with its Android platform powering roughly three-quarters of smartphones worldwide (StatCounter 2024), thereby supplying the firm with vast quantities of user data and further reinforcing its ecosystem. Unsurprisingly, Google's advertising services, anchored by its search engine and supported by extensive data collection and algorithmic targeting, captured a substantial share of

[192] Long before it developed its own browser, Google extended its reach through the popular Google Toolbar, an add-on for browsers like Internet Explorer. This toolbar embedded a Google search box directly into the user's browser, making search ubiquitous and funneling a massive volume of queries and browsing data back to the company (see Schmidt 2018).

global digital advertising revenues, clocking in at clocking in at roughly $235 billion in 2024, making it the world's largest digital advertising platform (Statista 2024b); indeed, that year its annual advertising revenue was an order of magnitude greater than that of major traditional media categories such as newspapers and radio, and substantially larger than global television advertising revenues (ibid).

To achieve this distinction, Google created openings for two distinct types of transactions across its network. The first one occurs on the search page itself and is based on a "Pay-Per-Click" model where the advertiser is only charged when a user actually clicks on their ad (Varian 2007). A second, separate, transaction can occur later as the user browses other websites in Google's vast ad network. Here, the advertiser might pay on a "Cost-Per-Mille" basis, which is a fee for every thousand times the ad is displayed to the targeted user (see UK CaMA 2020). In this second case, Google shares the revenue with the owner of the website where the ad appeared through its "AdSense" program (Google 2024).

With the stack assembled and distribution locked in through Android, Chrome, and Safari defaults, we can see the Google flywheel at work at the level of a single query. Imagine a user searching for "best running shoes for bad knees".[193] In that instant, the user provided a clear "signal of intent". After Google returned its organic, non-paid results, the user's click behavior acted as a "feedback signal" that helped Google's algorithm learn which pages were most

[193] This hypothetical scenario illustrates the interaction between (i) sponsored-search auctions and ad ranking, (ii) cookie-based tracking and behavioral advertising, and (iii) platform data collection and linkage across services. For background, see Varian (2007); Google (n.d.); FTC (n.d.c); FTC (2009); Grbovic et al. (2016); and Schmidt (2018).

relevant, improving the quality of future search results for everyone. In the milliseconds it took to load these results, a separate, lightning-fast ad auction took place where shoe companies bid on keywords related to “running shoes”. The winner’s ad was displayed at the top of the page. Moreover, as that user browsed YouTube or other websites in Google’s ad network, they might have seen targeted ads for those shoes, a practice known as “retargeting”. This single search query was thus transformed into a persistent data signal that increases the value of ad slots across the web. Google’s strategy in this case was predicated on the creation of a detailed user profile based on their browsing history, which was then aggregated with millions of others into highly specific “targetable audiences”.[194]

At scale, these billions of daily interactions create a self-correcting dataset that confers Google with a compounding advantage over its rivals across both data monetization and, as we shall see shortly, AI model training. In advertising, it moves the search engine beyond simple keyword matching toward correlation-driven personalization (UK CaMA 2020). Because every click serves as a human-validated "vote," the system statistically maps relationships between concepts to predict intent even for novel or obscure queries (Varian 2007). In LLM domain, this "conversation with humanity" (Battelle 2005) produces a curated repository of language evolution that is fundamentally superior to a static web scrape, allowing Google to leverage its trove of validated intent—combined with proprietary "crown jewels" like YouTube audio and the Google Books archive—to train models like Gemini.

Impact on the AI Revolution

[194] As already outlined earlier in this chapter, Google eventually shifted from individual-level tracking to more anonymized, cohort-based advertising.

Google's business practices and technological innovations both indirectly and directly helped bring about the AI revolution. While in the next chapter I will review the academic and technical advances behind the several decades long evolution of LLMs, culminating in the foundational models that dominated AI applications in 2025, here I explain that Google's pioneering approach to internet search and its targeted ad machine learning algorithms were just as instrumental.

Indirect Effects

Early on, Google's solutions to the fundamental problems of the digital age—finding relevance in a sea of information and monetizing user intent at scale—put in place the key technical and philosophical building blocks for generative AI. Its innovations in data collection, probabilistic algorithms, and content management created the blueprint that the AI revolution would later follow.

Consider that while a small number of queries are extremely common, most daily searches are unique, obscure, multi-word questions that have never been asked before. In fact, approximately 15 percent of all searches processed every day have never been seen by Google before (Gomes 2017). To answer these, Google's engineers moved beyond simple keyword matching to develop probabilistic algorithms that could infer the semantic intent of novel questions, which required training models on the entire web to learn the statistical relationships between concepts (see Domingos 2015; Nayak 2019). This pioneering work in understanding and responding to the "long tail" of human curiosity was a direct precursor to the core function of modern LLMs, which similarly use probabilistic models to generate novel responses to an infinite variety of prompts.

Filtering—or eligibility screening—first removes or down-weights content that violates policy or is judged to be low quality. Platforms rely on rule sets and machine-learning classifiers

that analyze text, images, and video to keep disallowed items from ever entering the candidate pool. Google's early implementation of SafeSearch and later AI-driven content moderation pioneered scalable ways to exclude spam, explicit content, or misinformation from its vast inventory (Schmidt 2018; see also Gillespie 2018). This pioneering work in scalable content moderation provided the blueprint for the essential safety filtering and data cleaning pipelines used to train modern AI models.

Once that pool is "clean," a ranking algorithm scores the remaining items for each query using signals like historical engagement, freshness, semantic relevance, and user behavior. The highest-scoring items win scarce on-screen slots, creating a feedback loop: user reactions to ranked content feed back into the ranker, refining its relevance models. From its inception with PageRank in 1998—a revolutionary algorithm that ranked pages based on link structures—Google transformed search into personalized, quality-driven results. This constant feedback loop, where user behavior refines algorithmic ranking, is the direct intellectual ancestor of the Reinforcement Learning from Human Feedback (RLHF) techniques used to improve LLMs (Christiano et al. 2017).

Google also transformed digital advertising from rule-based targeting to correlation-driven personalization based on predicting user intent (UK CaMA 2020). Its Ad Rank system, introduced in the early 2000s, combined bid amounts with quality scores—factoring in click-through rates, ad relevance, and landing page experience—to determine auction winners (Varian 2007). This system of using a quality score to predict and reward user engagement is a direct parallel to how modern AI models are fine-tuned to produce the most relevant and engaging responses.

Direct Effects

Moreover, the immense success of Google's integrated universe created what became perhaps the world's most valuable and dynamic dataset for training LLMs, giving the company a profound, built-in advantage in the race to commercialize AI. Unlike a static scrape of the web, this was a constantly updated, two-decade-long "conversation with humanity" (Battelle 2005). It captured "human intent" through trillions of search queries expressed in natural, conversational language. Importantly, it was also a curated and self-correcting repository. Every time a user clicked on a search result, they provided a high-quality feedback signal, implicitly telling Google: "this was a good answer to my question." This singular process, repeated billions of times a day, created a massive trove of real human questions paired with human-validated answers, all while capturing how language and priorities evolved in real time.

This strategic data advantage culminated in the development of Google's flagship AI, Gemini. The model was trained not just on the public web, but on a colossal, multi-modal dataset estimated to be in the trillions of tokens, drawing heavily from Google's proprietary crown jewels: the full archives of Google Search, the transcribed audio from YouTube, and the vast repository of Google Books (Gemini Team Google 2023; Metz et al. 2024). This allowed Google to build Gemini Ultra with an estimated 1.5 trillion parameters (Thompson 2024); as I will explain in the next chapter, each parameter is like a tunable knob that helps encode a nuanced understanding of language, logic, and concepts.

FACEBOOK'S ROAD TO AI RAN THROUGH ENGAGEMENT

Going into 2025, the core of Meta's business model consisted of maximizing user engagement—hosting and promoting content that users click on, share, and spend more time viewing. To that end, the company continuously refined its recommendation algorithms to personalize users' experience and keep billions of eyeballs glued to its platforms.

Meta's initial strategy, spanning from inception to about the late 2010s, was to stimulate intense, emotional forms of engagement within a user's existing social network.[195] At first, this was driven primarily by the binary "Like" button—a revolutionary tool introduced in 2009 that allowed the company to quantify user affinity and interest with a single click (see Kincaid 2009). Then, in 2015, Facebook introduced "Reactions", which let users deploy a wider range of emojis when responding to their friends' content, and thus allowed Facebook and Instagram to collect more nuanced data on users' sentiments and behavior (Krug 2016). To encourage people to relive and reshare past posts, Meta also began mining users' personal archives to spur new interactions, launching Facebook features like "On This Day" and adding topical prompts to their "Feed"—automated cues that remind users of past milestones or ask for their thoughts on trending topics—to stimulate conversations around current events (Gheller 2015).

Across these multiple engagement-increasing innovations, Meta patiently developed and honed its key 2010s era asset: the Social Graph, an established map of a user's real-world social connections and direct interactions (see Facebook 2007a). While Google's dataset was built on a web of information and user intent, Meta's was built on a rich, relational map of human society. Features like Groups, Pages, Events, and photo and video sharing are not just useful tools for

[195] Both for Meta and other digital platforms, the goal of maximizing user engagement sometimes incentivizes the algorithmic amplification of extreme or "borderline" content. Recommendation and ranking algorithms that optimize primarily for engagement metrics (e.g., clicks, shares, likes, and time spent) can learn that provocative, sensationalist, or emotionally charged material, including offensive, controversial, divisive, violent, or sexual content, is especially effective at grabbing human attention (Milli et al. 2025; Zubrow 2021).

users, but ways of capturing evolving data about real-world relationships, community affiliations, shared interests, and cultural trends (Facebook 2007a; Facebook 2007b; Facebook 2007c).

The resulting data churn powered Meta's monetization strategy, a sophisticated ad-targeting system built on the rich personal information users share across its platforms: their age, relationship status, the pages and activities they "like," the groups they join, and the content they interact with. When an advertiser creates a campaign, they can therefore select from a vast menu of parameters that goes far beyond basic demographics—and Meta's algorithms can even leverage a user's social context and friend networks to help fill in any gaps, all in service of creating a detailed profile that makes ad impressions more relevant and valuable (UK CaMA 2020; BigCommerce n.d.). This allows both Facebook and Instagram to target a recently engaged person between the ages of 20 and 30 who lives in New York City, likes swimming, and drives a BMW, for example.

To pull this off, Facebook engaged in very intensive tracking (Henry 2011). While logging out of that social network may have deauthorized a user's browser from the platform's web applications, it did not stop Facebook's cookies from sending information back to Meta's servers about where they went afterward and what they were doing there. Being that these insidious cookies never expired and were only altered, instead of deleted, when a user logged out, if they went on to visit another site with Facebook-enabled social applications, such as Like buttons, Meta could continue to track them.

While Meta engineers later modified their cookies so that users' account information was not present when they logged out of Facebook and Instagram, the company's most powerful tools continued to allow advertisers to leverage their own customer data through a secure

matching process known as hashing. At the time of this writing, a business can upload its own list of customer email addresses or phone numbers and securely match them to user profiles on Facebook and Instagram, allowing it to retarget existing customers with new offers. From this list of existing customers, an advertiser can then create a “Lookalike Audience,” a powerful feature that directs Meta's AI to analyze the characteristics of these best customers and find millions of new, similar users to target (Facebook 2022).

Another important driver of Meta's advertising dominance was its focus on creating a self-serve ecosystem that "democratized" access to sophisticated marketing tools, especially for small and medium-sized businesses (SMBs). While large corporations long employed specialized agencies, Meta made it easy and affordable for any local shop or online entrepreneur to become an advertiser. This strategy proved immensely successful; by 2017, Facebook had amassed over 5 million advertisers, the majority of whom were SMBs, with the fastest growth occurring in emerging markets outside the US (Facebook 2017).

To empower these millions of businesses, Meta built a comprehensive suite of user-friendly tools. The “Ads Manager” app served as a central dashboard to create campaigns, track performance with easy-to-read reports, and leverage “Advantage+ placements” (formerly automatic placements) that let Meta’s AI decide the most effective place to show an ad across Facebook, Instagram, and Messenger. To lower the creative barrier, resources like “Mobile Studio” taught entrepreneurs how to produce compelling video ads using only a smartphone (Facebook 2017). Meta also rolled out a massive educational effort, “Facebook Blueprint,”

which provided free online courses to millions of users on the fundamentals of digital marketing (Facebook 2015).[196]

Besides raking in ad dollars, Meta also cultivated a rich developer ecosystem by positioning itself as the internet's indispensable social infrastructure. Beginning in May 2007, the company gave tens of thousands of outside developers access to Facebook's "Social Graph" (see Facebook 2007). Several apps exploited this rich, interconnected web of friend connections, "likes," photos, and other user activities through the Graph Application Programming Interface (API).[197] For example, gaming giant Zynga built an empire with titles like "FarmVille" that were deeply woven into users' friend networks; dating apps like Tinder and Hinge used Facebook data to build profiles and suggest matches based on mutual friends; and music services like Spotify initially automatically shared users' listening activity to their timelines via the "Open Graph" protocol (Kirkpatrick 2010). Moreover, Meta maintained special data-sharing partnerships that granted preferential treatment to major partners While Microsoft's "Bing" was given the ability to see the names of virtually all a user's friends, Netflix and Spotify had permission to read, write, and delete users' private messages (Dance et al. 2018).

---

[196] Meta actively sought feedback to improve these tools by creating "Small Business Councils," and it developed specific products for its global audience, such as low bandwidth "Slideshow" ads for regions with slower internet and "International Lookalike Audiences" to help SMBs find new customers in other countries (Meta 2025).

[197] An API is a set of rules and tools that allows different software applications to communicate with each other. They define how requests for information or services should be made between systems, allowing developers to build on top of existing software platforms.

The New Era

In the years that followed, Meta's engagement strategy shifted to aggressively embracing new, more immersive content formats, often in direct response to emerging rivals. Following the rise of Snapchat, Meta rolled out the ephemeral "Stories" format —full-screen photos and videos that disappear after 24 hours and are pinned to the top of the interface—across both Instagram and Facebook, which encouraged more frequent, informal sharing (Hunter 2016; Hayes 2017). And on the heels of TikTok's explosive growth, Meta reengineered Facebook and Instagram entirely in 2022 to prioritize "Reels", an algorithmically driven, short-form video feed that allows users to scroll vertically through a continuous stream of personalized, high-tempo, recommended content created outside their friend group. This marked a pronounced move away from the "Social Graph" and toward a "Discovery Engine" that surfaces engaging content from creators across both platforms (Meta Platforms Inc. 2022; Malik 2022).

Implications for AI

Over the two decades leading up to 2025, Meta was not only able to create hyper-targeted ads and make lots of money doing so, but in the process also amassed an unparalleled dataset of natural, conversational, and social human interaction that gave it a competitive advantage in the AI race. While Google's data excelled at mapping human intent, Meta's "Social Graph" data allowed its models to learn the nuances of human social interaction, emotion, and culture. Meta's data corpus included the conversational language from public Facebook posts and comments, the specialized jargon from millions of Groups, and the rich multi-modal data from Instagram, where billions of images and videos are paired with user-generated captions, tags, and comments (Meta Platforms Inc. 2023).

While Meta engineers trained the company's "Llama" models on a massive dataset scaling to over 15 trillion tokens (Meta AI 2024) that required millions of GPU hours on Meta's custom-built AI supercomputers, the company chose to allow developers to freely access the models' weights and APIs, eschewing the proprietary, closed-off approach of its rivals OpenAI and Google. On the one hand, Meta's open-source strategy aims to commoditize the foundational model layer, preventing any single competitor from dominating the AI market. On the other hand, making the models free for the taking creates a "gravity well" that encourages the developer ecosystem to build on Meta-compatible infrastructure, potentially tying downstream applications to Meta's proprietary ecosystem and advertising APIs (see Zuckerberg 2024).

APPLE APPROACH AND CONTRIBUTIONS TO AI

Unlike Google and Meta, Apple built its unique multisided market by offering tightly coupled, high quality hardware and software features that attracted users to the iPhone first (Campbell-Kelley et al. 2015). This includes frictionless onboarding like the ability to seamlessly subscribe to apps with a single tap via In-App Purchase (IAP); location-aware features that make search faster and more relevant; interoperability between devices enabled by shared credentials through iCloud and synced preferences.[198] Then, once millions of eyeballs were captivated, advertisers, developers, and complementary hardware providers jumped on the Apple bandwagon, hoping to

[198] This interoperability is not merely a convenience but a strategic 'handshake' between devices—utilizing features like Handoff and Universal Clipboard to create a unified user experience that rivals cannot easily replicate without owning the underlying operating system (Apple Support n.d.).

gain access to a massive and loyal userbase (ibid).[199] While the massive reservoir of real-world data deposited downstream of this phenomenon would eventually prove invaluable for pushing AI progress forward, world-changing innovations that I outlined below emerged first.

Apple's Iconic Walled Garden

To turn the iPhone into a global powerhouse, Apple provided outside developers with a specialized "digital toolbox." It included Xcode and Swift (Apple Developer n.d.a). Most importantly, it featured a set of secure "gateways"—known as APIs—that allowed apps to safely use the iPhone's advanced hardware, such as its GPS, health sensors, and camera, without compromising users' privacy (Apple Developer n.d.b; Apple Developer n.d.c). By giving developers a standardized way to plug into features like Apple Pay (Apple Developer n.d.d), HealthKit, or Core Location, Apple ensured that the high-quality data generated by these apps was structured and secure (Apple 2024a). While these early tools focused on traditional app functions, Apple eventually supercharged this toolbox by integrating dedicated AI silicon directly into its devices—a hardware evolution that allowed the system to process complex intelligence tasks locally rather than relying on the cloud, as I detail in Chapter 10.

---

[199] Consider that, in 2003, Symbian had 56% of the early global proto-smartphone market; that share grew to an overwhelming 64% in 2007. When the iPhone launched in 2007, there were just 10,000 apps available for the Symbian platform, however—a number that had taken more than seven years to achieve. Sales of Symbian-based apps were soon dwarfed by Apple, which had 100,000 apps in its store within 18 months after launch. From its height in the year the iPhone was introduced, Symbian saw its share fall to 52% in 2008, 47% in 2009, 32% in 2010, 12% in 2011 to 1% in 2012, before disappearing completely (see Recon Analytics 2018).

A Synergy of Hardware and Software Innovation

On its own, the iPhone's hardware revolutionized several domains (see Richter 2023; Evans 2018). Mobile-first social media, including apps like Instagram, leveraged the iPhone's camera to create entirely new forms of visual communication. Additionally, the gaming industry was transformed from a console-dominated market into a mobile-first one, with franchises like Angry Birds and Candy Crush pioneering the "freemium" business model. In finance, apps like Venmo and Robinhood used the iPhone's secure platform to simplify everything from peer-to-peer payments to stock trading, democratizing access to financial services.

Then, when software developers architected GPS-enabled software, they looked to iPhone compatible hardware for help (see Porter and Heppelmann 2014). For example, health-monitoring apps call upon compatible Bluetooth accessories like smart scales and medical-grade oximeters that work seamlessly with iPhones. Similarly, smart home services for thermostat controls and lighting systems rely on iOS apps to integrate hardware, cloud services, and subscription upgrades. Snap-on game controllers turn an iPhone into a dedicated gaming console, and professional-grade external microphones make mobile podcasting possible.

Meanwhile, Apple's digital toolbox represents an operational template that any marketplace can adopt, allowing them to plug into iOS sensors and payments, orchestrate supply and demand, and scale nationally via the App Store (Apple 2023; Kenney and Zysman 2016). Take ridesharing as an example. "Core Location" (Apple Developer n.d.b) and "MapKit" provide real-time maps and routes that let a user see a ride approaching. "Push notifications" and "background services" (Apple Developer n.d.e) serve as the dispatch system, alerting a driver to a new job even if the app isn't open. Secure, one-tap payments are made possible by "Apple Pay", with the underlying financial data protected by the iPhone's "Secure Enclave" hardware

chip (Apple 2024a). Trust between strangers is also engineered through the platform: "Camera APIs" make driver license scans for identity verification possible, while in-app ratings and photo confirmations of deliveries represent an accountability layer (Apple 2024a).

Apple Monetized This Ecosystem

Apple was able to extract significant value from each side of its platform: by 2025, it captured a 15 to 30 percent commission on virtually every digital purchase or subscription that occurred inside the App Store (Leswing 2025).[200] Developers were required to use Apple's IAP system (Apple Developer n.d.d) for digital goods, ensuring the latter collected its cut of the revenues. While "real-world" services like ride-hailing or food delivery were exempt from the “Apple tax” on the transaction itself, the company took a cut from their high-margin subscription layers—

[200] While this tightly controlled arrangement was upended by the courts in a blockbuster antitrust ruling (see Mehta 2025), it proved very lucrative for Apple while it lasted. Consider that iPhone unit sales essentially plateaued in 2015. Apple sold roughly the same number of devices in 2024 as it did that year; by 2024, services made up nearly a quarter of its sales (generating over $96 billion) and close to 40% of gross profit, thanks to over 70% margins on offerings like iCloud storage, paid apps, in-app advertising, and search royalties, compared with Apple's overall margin of 32% (see Winkler and Rattner 2025). To put this in perspective: services accounted for under 10% of Apple’s revenue in 2015, contributing only about $12 billion in gross profit (Campbell-Kelly et al. 2015).

such as Uber One or DashPass—and through the specialized MFi licensing program for hardware accessories (Apple n.d.a).[201]

Finally, Apple's control over both device hardware and operating system allowed it to move from a passive host of third-party advertising to a dominant advertising juggernaut in its own right. By leveraging proprietary hardware signals—including GPS, Bluetooth, and Wi-Fi—Apple established high-fidelity "location APIs" that it used to certify third-party developer access while simultaneously building its own "Search Ads" business (Apple Developer n.d.f).[202]

This gave Apple the unique power to set the rules for behavioral tracking; while third-party networks like Meta's Audience Network were curtailed by Apple's privacy mandates, Apple's own "first-party" data allowed it to serve highly lucrative, intent-based ads within the App Store and across its native apps (Aridor et al. 2025). This pivot allowed Apple to distance itself from the opaque 'data broker' economy—where third-party trackers often harvest thousands

---

[201] Apple's "MFi" program is a licensing and certification process for third-party hardware manufacturers. In exchange for licensing fees, participating companies get access to Apple's proprietary technology, such as the "Lightning connector" and specific communication protocols, and the right to use the official MFi logo on their product packaging (Apple n.d.a).

[202] A location API is the specific tool that Apple provides for developers to securely ask the iPhone's operating system, "Where are we?" The API then delivers the device's GPS coordinates back to the app, enabling it to power features like turn-by-turn directions or show you nearby restaurants. Think of it as a secure messenger: the app can't access the sensitive GPS hardware directly; it instead must send a request through the official location API, which then provides the information in a standardized and privacy-controlled way (Apple Developer n.d.f).

of data points on a single user without their knowledge—by positioning its own ecosystem as a transparent, first-party alternative (Apple 2021a). Users accepted these prompts not merely for functionality, but because Apple successfully framed its "App Tracking Transparency" (Apple 2021b) as a privacy benefit, even as it redirected ad spending toward Apple's own walled garden (Leswing 2021; see also Chapter 10).

How This Fed into the AI Revolution

Apple's most significant contribution to the AI revolution was indirect. By creating the iPhone and the App Store, Apple built a vital infrastructural layer for the modern app economy (Kenney and Zysman 2016). Core to this was providing developers with secure, permission-based access to the iPhone's powerful sensors, especially its "Location Services," which allow apps to request real-time GPS, Wi-Fi, and Bluetooth signals only after explicit user authorization (Apple Developer n.d.b; Apple Developer n.d.c). This "permissioned access" regime helped enable a new generation of mobile-first digital platforms whose services were built around real-time, geo-tagged data, generating a torrent of behavioral information that existed outside of Apple's direct control.

By accelerating the growth of the gig economy, Apple's framework made data-intensive business models dependent on its technology (Kenney and Zysman 2016). Platforms like Uber and Lyft relied on the OS-level ability to provide continuous location updates to coordinate drivers and riders (Apple Developer n.d.e; Apple 2023). Food delivery services like DoorDash and restaurant review apps like Yelp created massive databases of culinary preferences and human movement tied to precise coordinates (Apple Developer n.d.f). Similarly, social media giants like Instagram—initially an iOS exclusive—thrived by tying their strategies to the iPhone's evolving camera and connectivity (Edelman 2015). Therefore, while Apple Intelligence

models are trained on responsible web crawling and licensed corpora rather than this "app exhaust," Apple's infrastructure created the conditions for this "digital commons" to come into existence (Apple 2024b; Apple 2025).

At the same time, Apple's tightly integrated ecosystem allowed the company to pursue a distinct data strategy. Positioning itself as a gatekeeper for an ecosystem of over a billion users, Apple's devices generate high-fidelity personal context that the company increasingly leverages for on-device personalization (Apple 2024b; Apple n.d.b). Unlike rivals that rely on the broad harvesting of the open web, Apple utilizes a vertically integrated stack to provide tailored experiences while emphasizing data minimization and privacy (Apple 2021c; Apple n.d.b).

Apple's approach to artificial intelligence therefore diverges from the cloud-centric models of OpenAI, Google, and Meta (Gemini Team 2023). Instead of centralizing private communications, Apple Intelligence is designed to provide personalization using on-device context—such as calendar events or frequently used apps—without Apple collecting the personal data itself (Apple n.d.c; Apple 2024c). This strategic pivot relies on a specialized hybrid infrastructure: simple tasks are handled by an on-device "Small Language Model," while more complex "Thinking" is shifted to a secure server environment (Apple 2024b; Apple 2024c; Apple 2024d). I detail this further in Chapter 10 by exploring how Apple built specialized AI hardware like the Apple Neural Engine and uses Private Cloud Compute to fulfill complex requests without ever retaining user content.

CONCLUSION

The AI revolution was not sparked by a single invention, but by a deliberate confluence of legal, commercial, and technological forces that transformed human interaction into a harvestable resource. As this chapter has shown, the "world-class pipes" built through market-based

telecommunications reforms provided the necessary physical foundation. However, it was the specific institutional architecture of the American internet—defined by the Section 230 liability shield and a Coasean approach to data ownership—that allowed digital platforms to scale without the friction of proportional legal risk or individual licensing.

By 2025, this ecosystem had produced a central paradox: while the open, user-generated web provided the initial fuel for AI progress, the benefits of that progress have largely accrued to the entrenched incumbents who control the "data spigots". Through two decades of search queries, social interactions, and mobile-first "app exhaust," Google, Meta, and Apple created a digital commons that they alone were uniquely positioned to refine. The resulting "data reserves" did more than just train models; they entrenched a new form of vertical integration where the race for supremacy favors those who control the entire stack—from the specialized silicon in users' pockets to the private cloud compute environments that process their most sensitive intent.

While this chapter has focused on the industrial organization of the data that feeds the revolution, the capacity of machines to actually "learn" from this information was not an overnight achievement. To understand how this vast digital commons was eventually transformed into reasoning agents and generative models, we must look at the industrial organization of the AI stack in its own right.

CHAPTER 9. INDUSTRIAL ORGANIZATION OF THE AI STACK

During the early 2020s, algorithmic breakthroughs in machine learning moved rapidly from research labs to the commercial market. Diverse applications evolved to automate complex tasks, leveraging computer vision to interpret the physical world, natural language processing to master human dialogue, and foundation models to generalize across domains without task-specific training.

Yet the ever-improving AI applications of the early 2020s—from coding assistants that transform natural language into software to autonomous agents that coordinate workflows—were only the visible tip of a vast technological iceberg. Behind every chatbot response and generated code snippet lies a "capital-intensive stack" of specialized providers and physical infrastructure that converts raw computational inputs into deployable intelligence.

To understand how this stack functions, consider the lifecycle of a single user request. When a user asks a question, an application layer first determines what the request requires, retrieves relevant context, and routes the task to an appropriate capability—whether a language model, a vision-language model, or a search tool. A foundation model then generates a draft response conditioned on that context, which the application post-processes by applying policy constraints and safety filters. Finally, monitoring systems capture signals about quality and reliability, creating a feedback loop that is essential for high-stakes enterprise deployment.

Each layer contributes distinct forms of value—and distinct sources of leverage. Infrastructure provides the raw scale and reliability required to run massive models. Foundation models provide the reusable competence that lowers the marginal cost of adding new tasks. Integration layers operationalize that competence, embedding raw intelligence into specific

workflows like legal review or medical diagnostics. Monitoring and governance layers support the trust necessary to integrate these systems into the regulated economy.

The political-economy implication is fundamental: value creation and bargaining power concentrate where substitutes are scarce, switching costs are high, and complementary assets are hard to replicate. This principle—familiar from the semiconductor and smartphone industries analyzed in earlier chapters—structures the entire AI stack.

To make sense of this assembly line, this chapter exploits this insight and shifts the lens from the internal logic of the algorithms to the industrial organization of building and operating them at scale. Much like the smartphone supply chain outlined in Chapter 5, relative prices coordinate the allocation of key inputs—data, compute, energy, networking, and software. These prices do not just determine costs; they shape strategic partnerships, harden switching costs, and ultimately dictate the industry's industrial organization. They explain why the market has moved toward tight vertical integration, with power concentrating where bottlenecks are most acute.

This chapter continues as follows. First, I map the AI supply chain from raw data to deployment, emphasizing the physical infrastructure beneath the software—chips, data centers, networking, storage, and power—and showing how tight coupling between adjacent layers creates chokepoints. Second, I analyze how those chokepoints reorganize bargaining power across the stack: when compute is scarce, switching costs are high, and distribution channels are proprietary, market structure matters as much as algorithmic ingenuity. Third, I trace the resulting organizational response—vertical integration and quasi-integration through long-term contracts, equity links, and “license-plus-talent” deals—as leading firms move upstream into silicon and infrastructure and downstream into devices and applications.

THE AI SUPPLY CHAIN

Figure 9.1 provides a high-level overview of the AI supply chain, depicting eight vertically arranged layers from raw inputs to deployed applications. At the top sit data sources—public web corpora, licensed feeds from data vendors, and user-generated content from social media, forums, and IoT sensors. This raw material flows into data collection, cleaning, and annotation, where it undergoes deduplication, labeling, and augmentation. A distinct data storage and movement layer manages the infrastructure for housing and transferring these massive datasets—from "cold" object storage for archival to "hot" memory for active training runs. The refined data then feeds into training hardware and physical infrastructure—AI chip clusters housed in specialized data centers with dedicated power, cooling, and networking. Software frameworks and Machine Learning Operations (MLOps) provide operational scaffolding. The foundation model development layer is where frontier labs train the large language and multimodal models that power downstream applications. Once trained, models move to the inference stack, where they are deployed through cloud platforms, vector databases, and edge devices to serve real-time requests. Finally, end users and applications—from consumer chatbots to enterprise integrations—are the terminal point downstream where value is ultimately captured.

Importantly, the dashed lines on the right side of the figure indicate that this is not a one-way flow: multiple feedback loops return signals upstream, creating continuous cycles of refinement that make user-facing distribution a strategic asset for improving model quality. Figure 9.1 also immediately reveals one of the reasons why this industry favors vertical integration. The feedback loops depicted on the right side of the figure illustrate how value creation at one layer depends on tight coordination with others: usage data improves training data, inference signals refine models, and deployment monitoring triggers retraining. Unlike the modular, horizontally specialized supply chains that characterized the smartphone era (Chapter

5) and the humanoid robot industry (discussed further below), the AI stack exhibits tight coupling between adjacent layers: data quality shapes model performance, model architecture determines hardware requirements, and inference optimization depends on end-to-end system design. These complementarities create high asset specificity—firms make co-specialized investments in data pipelines, training infrastructure, and deployment tooling that are costly to redeploy across partners—and they make outcomes difficult to contract for and verify at arm's length. The result is a structural pressure to internalize multiple layers of the stack rather than rely on market coordination across firm boundaries.

The sections that follow unpack each segment of the supply chain outlined in Figure 9.1, illuminating how AI evolved into highly valued services in such short order—and why vertical integration across the stack became increasingly pronounced. Three forces drove this consolidation: the contracting frictions created by asset specificity outlined above, double marginalization dynamics that made arm's-length pricing inefficient, and cross-layer complementarities that rewarded end-to-end coordination and optimization.

Figure 9.1 The AI Supply Chain.

[INSERT FIGURE 9.1 HERE]

Notes: This figure depicts the AI supply chain as eight vertically arranged layers, from upstream data sources to downstream end-user applications. Solid arrows indicate the primary flow of value creation. The three dashed curves on the right represent feedback loops operating at different timescales: the longest loop captures how usage data from deployed applications flows back to become future training data; the medium loop represents the CI/CD pipelines (automated processes for building, testing, and deploying updated models over an indefinite

lifecycle) and model-drift detection that trigger retraining through MLOps infrastructure; and the shortest loop reflects the rapid fine-tuning cycle in which inference-time signals (such as RLHF preference data) feed directly back into foundation model refinement. Representative firms and platforms are listed in italics beneath each layer. Many firms span multiple layers—a pattern analyzed in detail in the discussion of vertical integration that follows.
Source: author's compilation.

Data Sources

During the early 2020s, training an AI model meant feeding neural networks huge datasets comprised of billions of parameters. As Chapter 7 explained, engineers marshalled datasets with data from different sources—vast public corpora, proprietary feeds sold by data vendors, and the endless stream of user-generated content from social media, forums, enterprise systems, and IoT sensors—sometimes tailored for particular models.

Over time, these datasets grew rapidly in size. GPT-1 was trained on the BooksCorpus, a dataset containing over 11,000 unpublished books (Zhu et al. 2015). For GPT-2 in 2019, OpenAI created WebText, a 40 GB dataset scraped from 45 million high-quality outbound Reddit links (Radford et al. 2019). The scale then exploded for GPT-3 in 2020, which was trained by sampling 300 billion tokens from a massive mix of datasets that included a 570 GB filtered version of Common Crawl, the WebText corpus, and multiple digitized book collections (Brown et al. 2020).

Unlike neatly structured tables, much of this material arrived in free-form text, images, logs, or telemetry, capturing real-world behavior in all its complexity: GitHub comments revealed coding patterns, streaming-service logs exposed viewing habits, and geocoded IoT readings tracked machine performance across factories and cities.

Acquiring this data at scale required specialized infrastructure. Web crawling—where automated programs send billions of HTTP requests to gather raw HTML code—was the primary mechanism for harvesting public internet content. Organizations like Common Crawl ran these crawls continuously, publishing petabyte-scale web snapshots that served as foundational corpora for frontier models (Baack 2024).

While petabyte-scale web crawls provided the foundational corpora for early language models, a parallel market emerged during the early 2020s for bespoke training data that could not be harvested from the public internet. As frontier labs exhausted easily accessible text and confronted the limits of scale-driven improvement, they turned to a different source: human experts paid to generate task-specific datasets from scratch. The catalyst was RLHF, the training technique that gave ChatGPT its conversational fluency. RLHF required contractors to rate the quality of chatbot responses, and a second model trained on these ratings then rewarded the chatbot whenever it produced outputs that this "reward model" predicted humans would prefer. As models grew more sophisticated, so did the expertise required to judge them. Evaluating whether medical advice was sound required medical training; assessing legal analysis required legal training. The cheap, pennies-per-task crowdsourcing that had powered earlier generations of machine learning—labeling images of dogs and cats, drawing bounding boxes around stop signs—gave way to a market for highly credentialed professionals, a phenomenon I will elucidate shortly below.

Cloud marketplaces such as AWS Data Exchange and Snowflake Marketplace enabled enterprises to license curated datasets from vendors without running afoul of privacy or regulatory constraints. For proprietary data, companies built direct pipelines: API (Application Programming Interface—a set of rules and protocols that allows different software programs to

communicate with each other) integrations to pull social media feeds, sensor networks to capture IoT telemetry, and enterprise connectors to ingest internal documents and communications.

Data Collection, Cleaning and Annotation

After model trainers acquired data, this raw material remained far from ready for training. Duplicates abounded, quality varied wildly, and most content lacked the labels needed for supervised learning. Transforming this ore into refined fuel required a multi-stage refinement process—akin to panning for gold—starting with a vast riverbed of raw content and ending with a purified set of training-ready examples (see Provost and Fawcett 2013).

The first stage was parsing and extraction. Programming libraries like BeautifulSoup stripped away extraneous website code, ads, and navigation menus from crawled HTML, extracting only the clean, human-readable content. Similar extraction pipelines processed PDFs, converted images to standardized formats, and normalized raw operational data into structured records.

The subsequent curation phase was where the most detailed refinement occurred. Low-quality content was filtered out, near-identical pages were deduplicated, and harmful or biased text was removed. Data teams conducted spot checks and bias audits to ensure the dataset's quality and fairness. For supervised learning tasks, the data was then labeled—a labor-intensive process ranging from simple classification tags to complex annotations like bounding boxes for images. To bolster examples of rare but important scenarios—such as complex reasoning or coding problems—teams augmented datasets with synthetic data (see, e.g., Touvron et al. 2023;

OpenAI 2023).[203] Only after this rigorous process was the content ready for tokenization, where text was broken into sequences of words or sub-words.

The most labor-intensive products in this market were "grading rubrics"—detailed taxonomies specifying, in painstaking granularity, what counts as a correct or high-quality output in a given domain. A single rubric might require ten or more hours to produce and include dozens of criteria, and labs commissioned large volumes of such rubrics across major training and post-training cycles, often tailored to specific domains (Dzieza and Field 2025).

Alongside rubrics, vendors also built reinforcement learning environments—often simplified software "gyms"—where AI agents could practice multi-step tasks in simulated interfaces until they learned how to perform complex, multi-step tasks without risking costly real-world errors (Palazzolo and Muppidi 2025). Rather than relying solely on human demonstrations, these environments recreate applications like Salesforce, DoorDash, or Amazon in miniature, complete with dummy data and scripted user interactions. An AI agent can, for example, practice recording a new lead in Salesforce or placing an order on Amazon thousands of times, learning through trial and error which sequences of actions yield the highest "reward" signals, all within a safe, offline sandbox.[204]

---

[203] On the theory and effectiveness of this strategy see Nikolenko 2021; Jordon et al. 2022.

[204] Building effective RL gyms is challenging, as AI agents may discover unintended "shortcuts"—a phenomenon known as reward hacking. For example, an agent might learn to delete an entire database rather than fixing a single record because doing so technically maximizes its speed-based reward (Palazzolo and Muppidi 2025). To guard against this,

During the early 2020s, this data-processing ecosystem reportedly generated on the order of $10 billion in annual revenue (ibid) and developed distinct business models. Scale AI pursued volume, combining large teams of human reviewers with software platforms to organize annotations across millions of data points (Scale AI 2023). Surge AI pursued expertise, building networks of specialized contractors to create custom, high-quality datasets for frontier labs (Palazzolo and Weinberg 2025). Snorkel AI pursued automation, developing programmatic labeling methods that use developer-written heuristics to label data without manual review (Ratner et al. 2017). Weights & Biases pursued provenance, supplying infrastructure to version datasets—effectively a track-changes system—that helps make experiments reproducible regardless of sourcing method (Weights & Biases 2024). Together, these approaches navigated the tradeoff between scale, quality, and cost while maintaining the rigor required for frontier training.

The political economy of this market carried significant implications. First, it represented a form of value capture largely invisible in discussions of AI infrastructure dominated by data centers and semiconductors. In many cases, data vendors were among the few segments reliably turning profits—selling picks and shovels while frontier labs burned cash in pursuit of AGI. Second, the market's structure revealed something important about AI's trajectory. The premise of AGI is generalization: sufficiently capable systems should not require bespoke human data for each discrete task. Yet the opposite was happening. Labs were spending more on tailored data than ever, and improvements increasingly came from domain-specific post-training and

---

developers outfit RL playgrounds with carefully designed constraints and "guardrails" that penalize undesirable behaviors.

evaluation inputs. By 2025, high-quality human data had become a primary bottleneck for future progress (Dzieza and Field 2025). In this scenario, AI looked less like imminent superintelligence and more like a "normal technology"—potentially transformative, but requiring ongoing human input to extend capabilities into new domains, a topic I take up in the next chapter.

Data Storage and Movement

Before curated inputs are trained to create frontier AI models, there is one final step to consider.[205] To understand the hierarchy of AI data infrastructure, it is helpful to view the training facility as something like a high-performance commercial kitchen.[206] Cold storage acts as the massive supply warehouse located miles away—vast, inexpensive, and designed for shelf-life rather than speed. This is where the trillions of raw text tokens and images reside. Hot storage (typically NVMe flash) serves as the *mise en place* pantry located directly next to the line cooks: It holds pre-portioned, high-priority ingredients ready for immediate use.[207] This "hot"

[205] This subsection draws on Oladayo (2024).

[206] A similar analogy comparing data infrastructure to a restaurant kitchen can be found in Balp (2025).

[207] NVMe (Non-Volatile Memory Express) is a communication protocol designed specifically to unlock the speed of solid-state drives (Flash memory). Unlike older interfaces that were built for mechanical spinning hard drives and are limited to a single command queue, NVMe connects directly to the system's high-speed central bus and supports up to 64,000 parallel command queues (NVM Express 2024). This creates the equivalent of a multi-lane superhighway, allowing

tier must feed data to the GPUs (the chefs) without a millisecond of delay during the dinner rush, while simultaneously absorbing massive “checkpoints”—the equivalent of instantly freezing a perfectly cooked dish so the kitchen doesn't have to start from scratch if the power goes out.

In practice, this massive data reservoir is built on object storage technology, a market dominated by the major cloud providers (see Synergy Research Group 2024).[208] The industry standard is Amazon S3, with Google Cloud Storage and Microsoft Azure Blob Storage offering similar, highly durable, and relatively low-cost solutions for storing the petabytes of raw data required. On top of this foundation, companies increasingly adopt "lakehouse" architectures—pioneered by platforms like Databricks and Snowflake—that combine the massive, low-cost scale of a data lake with the structured query capabilities of a traditional data warehouse (Armbrust et al. 2021). These platforms provide tools for data scientists to catalog, clean, and prepare raw data directly, transforming the vast "junk drawer" of raw inputs into an organized "pantry" ready for training.[209]

---

the storage drive to handle tens of thousands of data requests simultaneously—a critical requirement for keeping power-hungry GPUs fed with data without pausing.

[208] Object storage is a data storage architecture designed to handle massive amounts of unstructured data. Unlike a traditional file system that organizes data in a hierarchy of folders (like on a PC), object storage places data in a flat, non-hierarchical "pool."

[209] I should emphasize that this "pantry" often requires a second, faster layer for the actual training run. While object storage is excellent for capacity, training clusters rely on parallel file systems to handle checkpointing—the process where a multi-day job periodically saves its progress. This operation is "bursty" because it creates a sudden, massive flood of data that must

Training Hardware and Physical Infrastructure

To train LLMs that grew in scale and complexity during the early 2020s, engineers distributed workloads across thousands of processors using model-parallel or pipeline-parallel architectures, with training runs extending over several weeks or even months. Several interconnected hardware innovations helped provide the massive computing power needed to pull this feat off. First and foremost, researchers relied on specialized processors, or GPUs, that excel at performing many simple calculations simultaneously, because neural network training involves millions of parallelizable matrix operations.[210]

A pivotal shift occurred with the dramatic densification of MACs—the fundamental circuits that multiply two numbers and add them to an accumulator. Deep learning is essentially a massive, parallelized sequence of these operations; training a neural network requires quadrillions of such calculations, making the efficiency of MAC hardware the primary bottleneck for AI speed. To accelerate this math, NVIDIA introduced Tensor Cores—specialized blocks built directly into the GPU that trade precision (using 8-bit or 16-bit math) for massive

---

be written instantly to prevent stalling the GPUs (Lee and Sengupta 2022). Parallel file systems are engineered specifically to allow thousands of GPUs to read and write simultaneously without creating traffic jams.

[210] Compared to CPUs, GPUs contain transistors laid out and connected in a way that allows them to do several matrix multiplications simultaneously; allow for larger batches of training data; and consume far less energy (NVIDIA 2024a). Therefore, GPUs were 10 to 100 times more cost-efficient and up to 1,000 times faster than CPUs for training models during this era (Dally 2023).

speedups in matrix algebra.[211] Google's TPUs took this concept further with Matrix Multiply Units (MXUs). These use systolic arrays—a highly efficient grid of simple processors where data is "pumped" through the chip in a synchronized rhythm, much like items on a factory assembly line, allowing for a continuous flow of calculations with minimal overhead (Jouppi et al. 2017).

Beyond optimizing individual units, chipmakers created larger and more complex processors by stacking multiple smaller chips (chiplets) using 2.5D and 3D integration. This approach replaces large "monolithic" dies with vertically integrated packages, offering three critical advantages. First, it increases manufacturing yields and reduces costs, as smaller silicon dies are less prone to defects than massive single chips. Second, high-bandwidth interconnects allow these chiplets to communicate with low latency, functioning effectively as a single, massive processor. Third, it enables "heterogeneous integration"—combining different

[211] Neural networks require quadrillions of MAC operations per second. To break this bottleneck, NVIDIA's "Tensor Cores" introduced a critical compromise: mixed precision. Instead of using traditional 32-bit numbers (which are like calculating with 10 decimal places of accuracy), Tensor Cores use 16-bit or 8-bit numbers. While technically less precise, this data type is "good enough" for model weights and allows the chip to process calculations exponentially faster—much like it is faster for a human to multiply two-digit numbers than ten-digit numbers (see NVIDIA 2024a).

components like CPUs, GPUs, and high-bandwidth memory in the same package to optimize every square millimeter for AI workloads (McMillan 2025; ASPDAC 2025).[212]

By 2025, custom-designed chips optimized for machine learning further advanced these developments, handling increasingly demanding tasks (Patterson and Hennessy 2021). Beyond traditional GPUs, innovative architectures emerged—Cerebras, for example, developed a "wafer-scale" processor, a single, massive chip built from an entire silicon wafer rather than being diced into smaller, individual processors—designed for significantly greater parallel processing and, according to the company, half the energy consumption of comparable Nvidia GPUs. Likewise, Google engineered custom TPUs optimized for both AI training and inference.

But raw processing power was not enough. To handle ever-growing workloads, engineers required optimized software frameworks, like NVIDIA's Compute Unified Device Architecture (CUDA), which I will discuss further below, and better networking—specifically, faster proprietary interconnects (such as NVLink) to facilitate sophisticated inter-GPU communication. This ensured that thousands of computing units could function as a single system, maintaining accuracy via mixed-precision calculations even while handling vast quantities of data (Dean et al. 2018; NVIDIA 2024a).

Supporting Infrastructure During Model Training

During the early 2020s, the tremendous heat and power demands of extensive computing clusters required advanced cooling and energy-management systems (Barroso et al. 2003). These clusters, often comprising thousands of GPUs, generate significant amounts of heat during

[212] BPD, which I flagged in Chapter 6, represents another significant development in AI chip design.

operation, which can damage equipment or reduce performance. Data centers had no choice but to invest in sophisticated techniques, such as liquid or immersion cooling, alongside traditional air conditioning. Simultaneously, operators adopted smart power management—such as automatically adjusting energy usage based on workload and sourcing renewables—to ensure that the massive cost of electricity did not make training large models economically impossible (see IEA 2025).

Given the exorbitant capital and technical requirements of building these powered "fortresses," few organizations could manage them in-house. This barrier to entry cemented the role of major cloud providers, who transformed the landscape by allowing organizations to treat supercomputing as a utility rather than a capital investment (Katsurashima et al. 2024; AWS 2024a). Instead of procuring physical hardware, teams could spin up specialized high-performance computing applications on demand, paying only for the capacity used (AWS 2024a; Microsoft 2024). These environments came pre-packaged with the essential plumbing for AI, including low-latency interconnects to minimize training bottlenecks (ibid). This shift allowed researchers to focus on innovation rather than infrastructure maintenance, drastically accelerating the iteration cycle (Katsurashima et al. 2024).

However, provisioning thousands of cloud GPUs solves only the computational half of the problem; to justify their cost, these processors must be kept fed. During a model's training stage, data storage functions like a Formula 1 pit crew—speed and precision are everything.[213] The expensive GPUs are the race car; if the pit crew is slow, the car sits idle and loses. This "pit

[213] Alvarez (2025) explicitly employs the Formula 1 pit crew analogy to explain asynchronous checkpointing in large-scale AI training.

crew" has two critical jobs: constantly streaming prepared data batches to keep GPUs running at full speed, and rapidly absorbing massive checkpoint files when the model saves its progress.

Finally, to ensure the pit crew never delays the car, the industry moved to a different class of technology: high-performance, all-flash parallel file systems. Led by companies like Pure Storage, VAST Data, and Dell Technologies, these systems use thousands of ultra-fast NVMe flash drives connected via proprietary networks like NVIDIA's InfiniBand.[214]

Together, this ecosystem of ancillary technologies—from sub-zero cooling and gigawatt-scale power management to ultra-fast storage—forms the essential scaffold that ensures that immense investments in GPU silicon are not wasted.

Nvidia's Enviable Pivotal Position in the Supply Chain

Leading up to 2025, the success of NVIDIA's AI-enabling GPUs was intrinsically linked to its strategic partnership with TSMC, a collaboration foundational to the modern AI stack. By leveraging TSMC's leading-edge process technologies—including advanced nodes in the 7nm/5nm/4nm/3nm families—NVIDIA manufactured increasingly dense and power-efficient AI

[214] VAST Data has gained significant traction by challenging the traditional tiered storage model with a fundamentally new architecture. Their core innovation, known as DASE or "Disaggregated, Shared-Everything," separates the storage system's "brains" (the servers that manage data) from the flash drives that hold it. Unlike traditional systems where processing power and storage are bundled together—like having to buy a whole new delivery van when you only need more cargo space—this disaggregated model allows companies to scale performance and capacity independently, much like adding a driver or a trailer to a fleet as needed (see McLaughlin et al. 2025).

accelerators for high-performance computing workloads (TSMC 2024). Indeed, NVIDIA's flagship platforms, including its much-vaunted Blackwell chips, which I will detail further below, were manufactured on custom TSMC processes (NVIDIA 2024b). This fabless arrangement allows NVIDIA to concentrate on GPU and systems design while relying on TSMC's scale, yields, and ramp capacity to meet booming global demand for its hardware (TSMC 2024; NVIDIA 2024b).

The company's hegemonic position reflected more than just raw silicon performance, however; Nvidia's CUDA software ecosystem consolidated a "walled garden" surrounded by a seemingly impregnable "moat" (see Mims 2024).[215] Consider that when developers train or finetune AI models they are not simply writing standard application logic; as explained in Chapter 7, they are architecting the massive linear algebra operations—primarily matrix multiplications—that underpin neural networks. To execute these operations efficiently, they must instruct the hardware on how to split billions of mathematical calculations into tiny chunks

[215] The characterization of Nvidia's ecosystem as a "moat" became the central, recurring metaphor in financial analysis during this period. By late 2025, a consensus emerged across major institutional desks—including Goldman Sachs, Morgan Stanley, and S&P Global—identifying "software-driven lock-in" as the primary justification for Nvidia's sustained 60%+ operating margins (see, e.g., Tsui and Chang 2025). This linguistic convergence was further evidenced by the launch of explicit "moat-breaking" initiatives, such as the UXL Foundation (hosted by the Linux Foundation), which united rivals like Intel, Google, and Qualcomm around a multi-vendor programming model designed to dismantle the proprietary barriers of the CUDA stack (iLinux Foundation 2023).

that run simultaneously. Frameworks like PyTorch, TensorFlow, and JAX serve as the high-level interface where developers design models using user-friendly code (Python). CUDA acts as the essential underlying "translator" that converts these high-level instructions into the precise, low-level operations of the GPU. Nvidia cemented its lead by building vast proprietary libraries of pre-optimized mathematical functions (such as cuDNN) that these frameworks rely on by default (Nvidia Developer 2025).[216]

To see why this creates such deep lock-in, consider that a programming model defines the abstraction layer between software and hardware.[217] Parallelism refers to how a task is broken into smaller sub-tasks that can be executed simultaneously by thousands of GPU cores. Mapping determines which parts of the code run on the host (CPU) versus the device (GPU) and how data moves between them. Kernels are the specific functions compiled to run on the GPU itself. Synchronization manages the timing between these thousands of parallel threads to ensure they don't overwrite each other's data or crash. Finally, portability refers to the ability of this code to run on different manufacturers' chips without rewriting it.

This final element is notably absent in the proprietary CUDA model. Because Nvidia has spent over a decade hand-optimizing specific mathematical blocks to match every transistor on their chips, they function as a performance multiplier. Competitors may offer raw chip speed, but

---

[216] The term cuDNN refers to CUDA Deep Neural Network library—a proprietary collection of highly tuned "primitives" (pre-written code blocks) for standard AI operations like convolution, attention, and normalization.

[217] On all these points see Biswas (2025).

without an equivalent library of mature, pre-optimized primitives, the same software often runs significantly slower or requires extensive manual retuning.[218]

This matters economically because interoperability shifts the demand curve. As Galetovic and Haber (2017) note, a platform becomes more valuable not just because its features improve, but because the universe of compatible "complementary goods"—in this case, software libraries and skilled developers—expands. In the CUDA ecosystem, this dynamic created a self-reinforcing loop of network effects. Direct network effects kicked in as a growing developer base shared code, expertise, and community support. Indirect network effects followed, as the massive installed base of Nvidia GPUs attracted the best frameworks, libraries, and tools. Over time, this success hardened into a trap: the accumulated weight of CUDA-specific code, performance tuning, and specialized human capital made migrating to alternative models prohibitively expensive.[219]

By ensuring that the richest set of complements—from framework optimizations (e.g., PyTorch) to hiring pools—worked best on Nvidia, the ecosystem endowed the company with market power that extended far beyond its silicon's performance. The gap between Nvidia and its closest challenger underscores this dynamic. Some estimates put NVIDIA's share of AI GPUs used in servers at close to 95% in 2025 (de Boer 2025). In fiscal 2025, Nvidia reported $115.2

---

[218] Recent benchmarks indicate that while rival hardware can match Nvidia's theoretical peak performance, real-world implementation often lags due to "missing kernels" and immature software support. This forces developers to spend valuable engineering cycles patching dependencies rather than training models (see Gupta 2025).

[219] See Farrell and Klemperer 2007 on the general point about network effects and lock-in.

billion in data center revenue and a GAAP operating margin above 60%, reflecting the extraordinary profitability of the AI-accelerator boom (NVIDIA 2025a). By contrast, AMD's data center segment generated only $12.6 billion in revenue in 2024 (see AMD 2025). Indeed, by the end of 2025, Nvidia's market value exceeded $4 trillion, making it the world's most-valuable company.

To better grasp Nvidia's hegemony in the AI chip space, Figure 9.2 plots the FP64 (double-precision floating point) performance (in FLOP/s) of flagship datacenter AI accelerators released between 2017 and 2025. Where the prior chapter tracked total training compute (FLOPs) used by frontier models, this figure tracks hardware throughput—the rate at which accelerators can perform floating-point operations. The dashed line represents an exponential fit to the data, implying an annual growth rate of approximately 1.3× per year—equivalent to a doubling time of roughly 2.7 years. While this pace is slower than the canonical Moore's Law trajectory of roughly 2× every two years that characterized semiconductor progress through the early 2010s (see Chapters 2 and 6), it is still quite impressive.

Several observations emerge from the data. First, NVIDIA's product progression—from the V100 (2017) through the GB200 (2025)—tracks the exponential trend line closely, confirming disciplined generational improvements in the 30–40 percent range. Second, AMD's Instinct series consistently exceeds NVIDIA on raw specifications. The MI300X, for example, delivers significantly higher peak throughput and more than double the memory capacity of its NVIDIA contemporary, the H100 (Epoch AI 2024). The fact that NVIDIA retains roughly 95 percent market share despite these superior "paper specs" (de Boer 2025) offers further empirical proof of the "walled garden" effect described above: for most buyers, the friction of leaving the CUDA ecosystem outweighed the raw horsepower advantage offered by the competition.

Looking beyond the horse race between these AI chip rivals, a key economic implication is clear. Because chip-level improvements in raw throughput lag far behind the explosive growth of frontier model training requirements—which often double every six months (Sevilla et al. 2022)—firms cannot rely on Moore's Law alone to solve their compute deficits. Instead, they are forced to close the gap through brute-force scale, tighter system integration, and massive capital expenditure. Indeed, the NVIDIA-TSMC partnership illustrates that even dominant hardware players depend on relationships across the stack. Yet silicon potential remains theoretical until it is aggregated into a coherent system. Hardware alone does not make AI work; the architecture for deploying that hardware at scale is equally critical.

Figure 9.2. AI Chip Performance, 2017-2025.

[INSERT FIGURE 9.2 HERE]

Notes: Data include flagship datacenter AI accelerators from NVIDIA (V100 SXM2, A100 SXM4, H100 SXM5, GH200, H200 SXM, B200, GB200) and AMD (MI250X, MI300A, MI300X, MI325X) released between June 2017 and February 2025. Performance is measured as peak FP64 (double-precision) throughput in floating-point operations per second (FLOP/s), without sparsity. While modern AI training increasingly relies on mixed-precision formats (e.g., TF32, BF16, FP8) that offer significantly higher effective throughput, FP64 is retained here as a consistent historical benchmark for raw silicon scaling across vendors and generations. Consumer GPUs, inference-optimized accelerators, and products from other manufacturers (Google TPUs, Intel Gaudi, Huawei Ascend) are excluded either because they do not report FP64 specifications or because they are optimized for lower-precision workloads not directly comparable on this metric. The trend line is fitted via ordinary least squares on $\log_{10}$-transformed performance values, yielding an estimated annual growth rate of 1.29×.

Source: Epoch AI (2024).

Furthermore, to defend its dominant position, Nvidia continued investing heavily in R&D and churned out several innovations in the process (see Leswing 2024). Its Blackwell chip featured larger die sizes (fusing two silicon dies into a single massive chip to maximize the surface area for transistors) and support for lower-precision arithmetic (using ultra-compact 4-bit numbers to speed up calculations and save energy), allowing it to boost throughput without a proportional spike in power consumption (Nvidia 2025b). Nvidia also deployed several Grace Hopper supercomputers, a hybrid architecture that bonds a CPU directly to a GPU to eliminate data transfer bottlenecks. And it pioneered new networking equipment to move information faster and drive down energy costs (InsideHPC 2024).

Inevitably, Nvidia's formidable supplier power triggered a countervailing reaction from the industry's largest buyers. In a deliberate bid to reduce reliance on Nvidia and erode its pricing leverage, OpenAI announced a landmark strategic partnership with AMD to deploy 6 gigawatts of compute capacity, beginning with AMD's upcoming MI450 generation in late 2026 (see AMD Inc. 2025). To grasp the sheer magnitude of this commitment, note that the industry now measures infrastructure not by counting individual chips, but by the power required to run them; six gigawatts is roughly equivalent to the output of six standard nuclear reactors.[220] By

[220] As AI clusters grew to include hundreds of thousands of GPUs, the industry shifted to using power capacity (megawatts and gigawatts) as the primary unit of measurement for infrastructure scale. This is because the physical bottleneck for these facilities is no longer floor space or server

dedicating this city-scale power capacity to AMD's MI450—the direct rival to Nvidia's frontier silicon—OpenAI effectively committed to building a parallel industrial ecosystem.

Indeed, the agreement was structured to incentivize this break from the CUDA monopoly: AMD issued OpenAI warrants for up to 160 million shares that vest only if specific milestones are met, including the successful scaling of non-Nvidia hardware By financially aligning OpenAI's success with AMD's stock price, the deal effectively subsidized the immense engineering effort required to optimize software roadmaps and overcome the switching costs that had long protected Nvidia's dominance. These moves paralleled industry groups' attempts to loosen CUDA's long-standing grip on AI developers by promoting open, cross-vendor programming models that enable code portability across different accelerators (see, e.g., Khronos Group 2025).

Software Frameworks and MLOps

We now turn to the intermediate layer of the AI supply chain: specialized software. Hardware provides raw computational horsepower; without orchestration it is inert, however. By coordinating data ingestion, training, evaluation, deployment, and monitoring, this software layer translates silicon capacity into usable systems.

At the foundation of this layer sit the development frameworks—PyTorch, TensorFlow, and JAX—that provide the high-level interface through which engineers design and train models. While these frameworks provide the design tools, they rely on NVIDIA's foundational library of pre-written mathematical operations to execute calculations on the chip; this default

---

rack availability, but the ability of the local utility grid to supply electricity (see Srivathsan et al. 2024).

reliance on proprietary libraries (such as cuDNN) reinforces the hardware lock-in described in the previous section (NVIDIA 2024a). PyTorch's flexibility made it the dominant choice in research settings, while TensorFlow's early investment in production tooling gave it an initial lead in enterprise deployment, and JAX attracted users seeking advanced optimization capabilities (Hu et al. 2025). By 2025, this fragmentation shaped which tools developers learned, as well as which workflows enterprises adopted.

To maintain models in the real world, by 2025 MLOps provided a systematic approach to updating models, auditing performance, executing automated rollbacks in the event of failure, and redeploying updated versions (Sculley et al. 2015). Specialized software empowered teams to collaborate on code, share pre-trained models, enforce version control, track experimental training runs, and detect "model drift"—the gradual degradation of performance as real-world data changes over time (e.g., Google Cloud 2024). To maintain efficiency, these tools leverage CI/CD (continuous integration/continuous deployment)—automated processes for building, testing, and deploying updates over the model's lifecycle (ibid).

This landscape of tools fits into four categories. First, open-source tools like Kubeflow manage pipeline orchestration and workflow control (Kubeflow 2024). Second, "all-in-one" commercial platforms—such as Databricks—offer unified environments that bundle storage, compute, and deployment (see Databricks 2024). Third, specialized trackers like Weights & Biases act as digital lab notebooks, recording the configuration and results of training runs (see Weights & Biases 2024). Finally, Hugging Face serves as the industry's shared public library, a central hub where developers collaborate on pre-trained models and datasets (see Jiang et al. 2023). Together, these tools reduce coordination costs but also embed workflows, making the software layer a durable source of switching costs in enterprise AI.

Foundation Model Development

At this layer of the supply chain sit the frontier laboratories—OpenAI, Anthropic, Google DeepMind, Meta AI, and Mistral, among others—that build the large pre-trained models powering downstream applications.

Foundation model development is extraordinarily capital-intensive (see Sevilla et al. 2022). By 2025, a single frontier training run could cost hundreds of millions of dollars in compute alone. Furthermore, the procurement of high-quality training data—the raw feedstock of the industry—became a parallel bottleneck, forcing labs to negotiate complex licensing deals with publishers as the supply of public web data tightened (Villalobos et al. 2022). These barriers to entry concentrated model development among a handful of well-funded players.

These capital and resource requirements reshaped corporate finance in the AI industry. As I discuss in the vertical integration section below, frontier labs increasingly relied on hyperscaler commitments—Microsoft's multi-billion-dollar partnership with OpenAI, and Amazon and Google's multi-billion-dollar investments in Anthropic—creating quasi-integrated relationships that blurred the line between supplier and customer (see Microsoft 2025a; Amazon Staff 2024; Field 2025a). Hybrid arrangements of this sort granted AI labs access to scarce compute while giving hyperscalers privileged access to frontier models—a mutual dependence central to the industry's evolving structure.

These strategic choices also produced distinct business models at the foundation layer. Closed-source leaders like OpenAI and Anthropic monetized through API access fees, subscriptions, and enterprise licensing, retaining strict control over model weights and architecture to protect intellectual property and mitigate safety risks (OpenAI n.d.). Open-weight providers like Meta pursued a different logic: by releasing Llama under permissive licenses,

Meta sacrificed direct model revenue but drove adoption of its ecosystem, positioning itself to commoditize the complement (foundation model weights) and capture value at adjacent layers such as advertising (Touvron et al. 2023). Hybrid players like Mistral occupied the middle ground, releasing high-performance open weights to build developer goodwill while selling proprietary larger models and managed services to enterprise clients (Mistral AI 2024).

While the dominant narrative of the early 2020s held that only hundred-billion-dollar clusters could produce frontier intelligence, the emergence of DeepSeek in late 2024 and early 2025 offered a sharp counterexample. The Chinese AI lab released DeepSeek-V3 and the reasoning model DeepSeek-R1, achieving performance comparable to leading Western systems on several benchmarks while emphasizing unusually low reported compute costs. DeepSeek's innovation was twofold: it optimized the base model using a Mixture-of-Experts design that activated only a fraction of parameters per token, and it pioneered Group Relative Policy Optimization (GRPO)—a reinforcement learning technique that incentivized the model to verify its own reasoning without the massive computational cost of traditional training methods (Guo et al. 2025; Xu 2025).[221] This development seemed to shift the economic bottleneck from the massive capital expenditure of training to the operational expenditure of running the models, partially democratizing access to frontier capabilities (see next chapter).

Even though DeepSeek successfully replicated the performance of existing models (like GPT-4o) at a lower cost, the frontier continued to move. As industry analysts noted, the release

[221] DeepSeek also benefited from a window of opportunity: before the US government tightened export controls on AI chips headed to China, the company had acquired 2,048 Nvidia H800 chips, which were compliant with earlier restrictions (see Woo and Huang 2025).

of next-generation systems like OpenAI's o3 delivered such massive leaps in capability that they simply could not be matched by clever engineering alone (Patel et al. 2025).

Inference Stack and Cloud Deployment

Once an AI model is trained, it is deployed to an environment where it can serve real-time requests using sophisticated tools. Specialized chips may have to power real-time image processing, complex language understanding, or large-scale recommendation engines (see, e.g., NVIDIA 2025b). Moreover, these computationally demanding AI model inference tasks might call on a separate cloud infrastructure optimized for inference (AWS 2024b; Google Cloud 2025a). By 2025, the economics of inference had bifurcated into two distinct workloads, each demanding different hardware strategies.

First, there was the traditional demand for high-efficiency inference: serving millions of simple user queries (like "write an email" or "summarize this text") as cheaply and quickly as possible. This favored chips that prioritized low latency and energy efficiency over raw power. For high-usage deployments, the aggregate cost of serving inference queries often exceeded the cost of the original training run (Aminabadi et al. 2022). NVIDIA responded with its Blackwell architecture—the details of which I explained above—explicitly engineered to handle both workloads. (Nvidia 2024b).

Second, the rise of AI reasoning models (like DeepSeek-R1 and OpenAI o1) shifted inference from a low-cost utility into a massive computational workload that rivaled training in its intensity. Unlike earlier models that produced direct answers in one pass, reasoning models deliberate internally---generating and checking intermediate steps before responding---so they can solve harder problems more reliably. These models require the chip to “think” for seconds or even minutes—allocating substantially more inference-time compute to generate intermediate

reasoning traces—before producing a final answer. To sustain these long "chains of thought," engineers had to overcome a new bottleneck: memory capacity. Because reasoning models must remember and attend to thousands of self-generated tokens, the industry raced to deploy chips with massive High Bandwidth Memory (see, e.g., Nvidia 2024a) and developed software techniques to manage the exploding temporary memory (KV cache) required to store long-context attention during inference (ibid).

The Rise of Alternative Architectures

However, relying on general-purpose GPUs to sustain these massive 'thinking' workloads proved economically daunting, fracturing the market into a landscape of three distinct architectural approaches. Companies like Groq and Etched developed Application-Specific Integrated Circuits (ASICs) that abandoned GPU versatility entirely. By baking the transformer architecture directly into the silicon, they achieved near-instantaneous token generation for real-time agents, though at the cost of flexibility (Mims 2025a). Hyperscalers similarly deployed custom ASICs—Amazon's Inferentia, Google's TPU v6, and Microsoft's Maia—to drive down the cost of serving their own internal models (Gardizy and Liu 2025). Seeking a middle ground between efficiency and flexibility, firms like Positron utilized Field-Programmable Gate Arrays—circuits that can be rewired after manufacturing. This allowed them to update the chip's logic to match the latest model architectures (like Mixture-of-Experts) without waiting years for a new physical chip cycle, drastically cutting power consumption per query (see Patterson and Hennessy 2021). Players like SambaNova introduced Reconfigurable Dataflow Units (RDUs), which physically map the neural network onto the chip to minimize data movement, offering a distinct advantage for the massive "thinking" workloads required by reasoning models (Prabhakar et al. 2024).

This strategic need to manage costs and diversify compute partners emerged as a central theme in the industry. Frontier model developers, wary of being locked into a single ecosystem, began to pursue multi-cloud strategies. A prime example is Anthropic, which secured billions in compute commitments from both Amazon AWS and Google Cloud (Anthropic 2025c). This dual-cloud approach allowed AI labs to leverage competition between the hyperscalers to secure scarce GPU capacity while accessing different custom accelerator architectures for their distinct inference workloads (Wiggers 2024a).

Cloud Infrastructure: The Global AI Nervous System

If model training is the "factory" where intelligence is forged, the cloud functions as the global logistics network that delivers it to the end user. By 2025, the primary value proposition of cloud providers (AWS, Azure, Google Cloud, Oracle) for inference had shifted from simple server rental to so-called elastic orchestration (Katsurashima et al. 2024). Unlike training workloads, which run at 100% utilization for months, inference traffic is highly volatile—spiking during business hours and dropping at night. Cloud providers managed this "burstiness" through serverless inference, allowing companies to spin up thousands of GPUs instantly during a viral moment and scale down to zero when traffic subsided, paying only for the milliseconds of compute that was actually used (e.g., AWS 2024b).

In a sense, the cloud became the abstraction layer that hid the hardware complexity described above. As the chip market fractured into general-purpose GPUs, specialized ASICs, and reasoning-optimized RDUs, cloud providers positioned themselves as the universal interface. Their "Model-as-a-Service" platforms (like AWS Bedrock or Azure AI Studio) acted as intelligent routers, automatically directing a user's prompt to the most economic hardware available—sending a simple summarization task to a cheap CPU or TPU, while routing a

complex coding problem to a powerful H100 cluster (AWS 2025; Azure AI Studio 2025). This flexible orchestration allowed enterprises to harness the efficiency of the new hardware landscape without needing to rewrite their software for every new chip architecture (Katsurashima et al. 2024).

However, access to compute is only half the inference equation. By 2025, for a model to be useful in a corporate setting, it had to be able to access the right information at the right time.[222] Once an AI model was deployed, its storage requirements shift from the massive datasets used in training to the precise, rapid-access systems needed for Retrieval-Augmented Generation (RAG).

To understand this architecture, consider a "Librarian Analogy."[223] Think of the core AI model as a brilliant librarian who has read every book in existence—but has been locked in a windowless room for two years. They have perfect memory of history, but no knowledge of current events or your company's private files.

The RAG system acts as their personal reference desk. In this setup, a vector database functions like a set of hyper-organized index cards, allowing the system to instantly locate the exact page of a relevant document based on meaning rather than just keywords. Meanwhile, a local document store keeps the actual reference materials within arm's reach. For speed, this entire reference desk must sit physically close to the inference GPUs, often using high-speed caches to prevent the "librarian" from stalling while waiting for information.

---

[222] This subsection draws on Ma et al. 2021.

[223] Similar "librarian" metaphors have been used by Wong (2024) and Ray (2024) to illustrate RAG's role.

A new ecosystem of companies emerged to provide these "index card" systems. Startups like Pinecone, Weaviate, and Milvus built dedicated vector databases, while established giants like MongoDB and Redis integrated vector search into their platforms (Zaitsev 2024).

Edge Computing and Small Language Models

Finally, Edge Computing represents a distinct configuration of the AI supply chain where processing occurs directly on devices—e.g., smartphones, laptops, and cars—rather than in the cloud. By 2025, this layer was defined by the rise of Small Language Models (SLMs): compact but capable AI models (such as Google's Gemini Nano) designed specifically to run within the power and thermal limits of consumer hardware (Gemini Team 2023; see also Apple 2024d; Apple 2025).

The "inference at the edge" shift, which was about optimizing the Apple Neural Engine to handle complex Transformer-based models with high efficiency and low latency (see Apple 2022), was epitomized by Apple Intelligence, released in 2024. Apple introduced a hybrid architecture that routed simple requests to an on-device foundation model consisting of approximately 3 billion parameters (Apple 2024b), while sending complex queries to Private Cloud Compute (PCC)—a dedicated server fleet built with Apple silicon that deletes user data immediately after processing (Apple 2024c; Apple 2024d).[224]

[224] Apple's "on-device first" logic is sustained by a sophisticated set of developer tools that allow apps to feed high-fidelity context into these models without compromising the system's security. By using APIs for "Core Location" and "background services," developers can ensure that the AI has access to real-time user intent—such as physical movement or proximity—while the

Apple's approach offered two strategic advantages. First, privacy: personal context (messages, calendar data) remains on-device for routine tasks, with cloud requests processed ephemerally and not stored (Apple n.d.b; Apple 2024c). Second, economics: by offloading the vast majority of routine queries to the user's own battery-powered device, Apple reduces the massive computational overhead and latency associated with cloud inference (Apple 2022; Apple 2024b; Apple 2024c).[225]

This dynamic helped shift the balance of power in the supply chain back toward device manufacturers who control the integrated stack of proprietary silicon and software (Apple 2024a; Apple 2022). Indeed, AI inference at the edge was migrating into wearables in the form of smart glasses, health-monitoring rings, and AI-enhanced earbuds. Along these lines, I now turn to exploring the touchpoints where computational infrastructure translates into economic value—actual goods and services that consumers and businesses are willing to pay for.

End Users and Applications

During the early 2020s, the AI supply chain culminated in consumer-facing and enterprise applications—the most visible being generative AI. Chatbots like OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini replaced complex code with simple conversation, allowing anyone to query massive intelligence systems using plain English (see Vaswani et al. 2017). Similarly, image and video tools like DALL-E, Midjourney, and Runway transformed

---

underlying data remains shielded by the iPhone's "Secure Enclave" (Apple 2024a; Apple Developer n.d.b; Apple Developer n.d.e).

[225] On-device inference can also cut energy use by up to 95% and water consumption by 96% compared to cloud processing (Sukumar 2025).

creative work by using "diffusion" models. These systems, having learned the relationship between words and pixels from billions of examples, could instantly conjure photorealistic scenes from a simple text prompt (see Ramesh et al. 2021). While businesses plugged these engines directly into their own software, millions of individuals simply downloaded an app, effectively putting a supercomputer in their pocket.

AI Agents

Perhaps the defining application-layer development of 2025 was the transition from chatbots to agents—AI systems capable of taking autonomous action, not just answering questions. These tools are designed to handle real work in business and consumer settings such as automating software development, executing web-based workflows, and conducting deep research.

Agents operated along a spectrum of autonomy. At one end were “Copilots” that worked alongside humans in real time, offering suggestions while keeping the user in control. GitHub Copilot, Amazon CodeWhisperer, and similar tools embedded AI directly into developers’ workflows, proposing code completions and debugging assistance. At the other end were fully autonomous agents capable of chaining multi-step tasks—planning, executing, and iterating without human intervention (see Efrati and Palazzolo 2025).

Coding agents exemplified this shift. By late 2025, tools like Devin, Codex, and Claude Code could autonomously write, test, and debug entire software modules. OpenAI reported that Codex alone was responsible for over two million merged pull requests on GitHub—a milestone signifying that human maintainers had reviewed the AI's work and deemed it robust enough to be integrated into live software (OpenAI 2025e). These agents didn’t merely autocomplete code; they could interpret high-level instructions, decompose them into subtasks, execute against live codebases, and recover from errors.

By late 2025, the coding agent market had already begun to consolidate. The top three players—GitHub Copilot, Anthropic's Claude Code, and Anysphere's Cursor—captured over 70 percent of a $4 billion market, with switching costs rising organically as developers adapted their workflows, prompts, and codebases to specific tools (CB Insights 2025a). Leading players moved to reinforce their positions through acquisition: Anthropic purchased Bun, a JavaScript runtime, while Anysphere acquired Graphite, an AI code review platform (ibid).

Browser-based agents represented a second major category. Anthropic launched Claude for Chrome, a sidebar assistant that could physically take over the browser—clicking links, scrolling pages, and typing into text fields on the user's behalf. OpenAI introduced its Computer Using Agent to handle web-based chores, while Perplexity released Comet, a browser built entirely around AI navigation (see FillApp 2025). These agents transformed the internet from a library into a personal concierge: a user could simply ask to "book a flight to Denver under $400," and the agent would autonomously scour travel sites, compare options, and fill out the checkout forms, eliminating the need for a human to manage every click and keystroke. Completing such transactions autonomously, however, required new payment infrastructure—prompting established card networks like Visa, Mastercard, and American Express to invest in startups building rails for agent-to-agent commerce (Hamid 2025).

Research agents constituted a third category. Google's Gemini Deep Research and similar tools could synthesize information across hundreds of sources, producing comprehensive reports on topics ranging from due diligence to drug toxicity safety research (Bort 2025a). Unlike traditional search, these agents didn't merely retrieve documents; they read, compared, synthesized, and structured information into coherent analyses. Crucially, these systems operated recursively: detecting gaps in their own initial findings and autonomously launching follow-up

searches to resolve ambiguities—automating the iterative investigative work previously reserved for junior analysts (see Google 2025b).

Finally, while in the introduction to this book I discussed some of the shortfalls that bedeviled AI agents in 2025, in the next chapter I will discuss how their potential may yet be realized as major foundation model developers collaborate on improving their underlying infrastructure. Rather than pursuing proprietary lock-in, OpenAI, Anthropic, and Block co-founded the Agentic AI Foundation in late 2025 to craft a set of agent interoperability standards (see OpenAI 2025e). As we shall see in Chapter 10, this parallels similar efforts during previous industrial revolutions that proved pivotal for consolidating and spreading new general-purpose technologies as documented in Chapter 2.

Enterprise Integration and Specialized Applications

During the early 2020s, several AI applications emerged to serve industries where regulatory compliance and domain expertise created high barriers. In finance, H2O.ai provided automated machine learning frameworks optimized for trading strategies, risk assessment, fraud prevention, and compliance reporting (H2O.ai 2025). In healthcare, Tempus aggregated clinical notes and genetic sequencing data to match cancer patients with specific therapies and clinical trials, ensuring that treatment decisions adhered to rigorous medical protocols (Tempus 2025). For the legal sphere, Luminance deployed its "Legal-Grade" AI platform to automate contract review, due diligence, and obligations tracking across enterprises (Luminance 2025).

By pre-packaging domain expertise and bespoke workflow integrations that general-purpose models could not provide, these specialized applications also raised switching costs that insulated vendors from commoditization dynamics. Indeed, industry-specific agent startups surged over 2025—the number targeting healthcare and life sciences alone grew from 7 to 47 in

just eight months—reflecting a bet that highly regulated industries would favor specialized solutions over horizontal tools (CB Insights 2025b).

AI-powered platforms similarly democratized data analysis. Tools like Databricks and Snowflake allowed analysts to bypass complex coding entirely; instead of writing SQL scripts, a user could simply ask, "Show me sales growth by region," and the system would instantly generate the necessary code and visualization (Databricks 2024). This shift extended to enterprise operations: Customer Relationship Management systems began automatically flagging clients at risk of cancelling before they called to quit, while Human Resource platforms scanned thousands of resumes to surface the best candidates—even identifying current employees whose skills matched open roles (Dadaboyev et al. 2025).

Finally, low-code platforms democratized software creation, allowing anyone to build applications simply by describing them. Tools like Microsoft Power Apps Copilot and Retool AI enabled a manager to say, "create an app to track warehouse inventory," and the system would instantly generate the screens, logic, and data tables required to make it work (Microsoft 2025b). By turning natural language into functional software, these platforms compressed weeks of engineering time into hours, transforming non-technical employees into "citizen developers" who became deeply integrated into the vendor's ecosystem.

In light of all of this progress on multiple fronts in the AI agent space, enterprise software incumbents moved quickly to capture growing opportunity, transforming their platforms from passive “systems of record” into active “systems of action” (Goode 2025). Salesforce launched Agentforce, embedding agents that could independently resolve customer shipping disputes or qualify sales leads without human oversight (Lessard 2025). Similarly, ServiceNow introduced

agents that could proactively diagnose server outages and reroute IT traffic before a human engineer even logged in (Zilbershot 2025).

Furthermore, incumbents held a decisive advantage in the AI agent race: their agents sat directly on top of the enterprise's historical data, allowing them to make high-stakes decisions with deep context that third-party startups—dependent on slow, restricted APIs—could not match (Yuan and Verçosa 2025).

Embodied AI and Humanoid Robotics

Over the early 2020s the AI supply chain increasingly extended beyond screen-based interfaces, into the physical world. Beyond smartphones and IoT sensors, AI inference also migrated into wearable form factors—smart glasses like Meta's Ray-Ban collaboration, health-monitoring rings, and AI-enhanced earbuds—representing further distribution of intelligence to the network edge. While embodied applications drew on the same foundation models, inference stacks, and training infrastructure described above, they added new layers of complexity: real-time sensor fusion, physical manipulation, and safety-critical control systems. A robot cannot just generate text; it must fuse blindingly fast inputs from cameras and lasers to "see" its environment, physically grasp objects without crushing them, and make split-second safety decisions where a software bug doesn't just crash an app—it crashes a car.

Amazon proved itself to be a clear leader in this field, deploying machines that mastered these very challenges. Inside its fulfillment centers, the Sequoia system demonstrated precise physical manipulation, using computer vision to identify and grasp specific inventory bins from millions of options, reducing order processing time by 25% (see Fielding 2023). Moving beyond the warehouse, Amazon leveraged this expertise to solve the even harder problem of safety-critical control in unconstrained environments: By late 2025, the company was testing humanoid

androids—including the agile Unitree G1—on rigorous indoor obstacle courses. These robots relied on advanced sensor fusion to navigate unpredictable terrain and dodge hazards in real-time, validating the "motion planning" algorithms required for autonomous final-mile delivery (see Drew 2025).

As they raced to deploy AI-powered robots in manufacturing, logistics, and eventually domestic settings, a wave of competitors—including Figure AI, Boston Dynamics' Electric Atlas, and Agility Robotics' Digit—attracted a surge of investment during that period, with over $21 billion in 2025 alone (Aggarwal and Mulé 2025). Tesla, however, bet the most aggressively. CEO Elon Musk claimed that humanoid robotics could eventually account for 80 percent of Tesla's value (Levy 2025)—a projection that, if realized, would dwarf the company's automotive business and could signal a fundamental shift in the economy: the potential decoupling of labor from human population growth (Jonas et al. 2024).

Accounting for Each Segment's Contribution to the Final Service's Value

Any attempt to slice the AI supply chain into neat value-added wedges—to precisely divide and measure the economic contribution of each part of the AI development and deployment process—remains elusive. Researchers cannot easily determine what portion of the final price should be attributed to each input's marginal contribution to the total economic value created when something like an AI chatbot is exchanged between sellers and buyers. For several interconnected reasons, trying to figure out the exact value added at each stage is far more complicated than when a traditional manufacturing process transforms inputs into a tangible finished good.

Unlike a product made on an assembly line, the cost and revenue data for the steps that went into creating a commercial AI service circa 2025 were deeply opaque. This information

was often scattered across proprietary contracts made with companies like Amazon, Google, or Microsoft for their exclusive cloud computing services, bundled platform fees, enterprise support agreements, and various negotiated discounts (PwC 2025). And while some of these costs were indeed measurable—for example, per-token API fees based on small units of text were sometimes less than $0.01 per thousand tokens when used at a large scale (OpenAI 2025f)—they varied widely, making it nearly impossible to disaggregate value: data-acquisition and annotation costs varied widely—from commodity tagging to expensive domain-expert labeling.

Even when list prices are public, they rarely map clearly onto marginal cost. API pricing is complex—often tiered, heavily discounted for volume, and bundled with "hidden" premiums like Service Level Agreements—contractual guarantees that the model will remain online and fast, even during traffic spikes (Machado 2025). This bundling is strategic. By packaging raw compute with these reliability promises and support services, providers can recover fixed costs: screening customers by willingness to pay, thus allowing them to extract higher margins from wealthy enterprise clients while filtering out casual users (see Tirole 1988).

Putting the opacity issue aside, companies populating the AI supply chain circa 2025 had to engage in steep fixed cost recoveries in part through their pricing strategies. As will become clear ahead, building datacenters required enormous upfront investments, sometimes running into hundreds of millions of dollars; for example, expenses include purchasing or renting thousands of NVIDIA's high-end H100 GPUs, which—depending on the specific configuration and purchase contract—cost between $25,000 and $40,000 each (Adeboye 2025).

And then there was the fact that the distinction between marginal and fixed costs was quite blurry to start with. Many organizations retrained or fine-tuned their models every few weeks or months, adjusting pre-trained models with smaller, specific datasets to improve their

performance for particularized tasks. This effectively turned six- and seven-figure CapEx hits into quasi-operational expenses over time. The same was true for personnel costs: staffing levels for roles like data engineers or labelers did not change proportionally with model usage.

Value attribution is further complicated because many AI products are implicitly two-sided: providers price not just the model output, but access to distribution, user feedback, and data exhaust.[226] In consumer settings, the "price" of inference may be partly recouped through advertising, bundling with other subscriptions, or lock-in to adjacent services. In enterprise settings, the model can be priced as a complement to workflow integration and risk reduction rather than as compute consumption (Levin and Ruiz-Reyes 2025). In both cases, observed prices reflect strategic ecosystem positioning as much as technical value.

In this environment, the central economic question is not "who adds how much value," but "who controls the irreversible complements"—chips, power, proprietary data, and distribution—that determine whether value can be realized at scale.

Where the Bargaining Power Lies

The AI supply chain described above is not merely a technical pipeline; it is also a bargaining system in which relative prices coordinate activity, but power is exercised through bottlenecks, switching costs, and control over complementary assets. As with other complex supply chains,

[226] In Chapter 7, I outlined the peculiar economics of two-sided markets. While historically "data exhaust" referred to the passive clicks and location signals used for ad targeting (see Zuboff 2019), in the generative AI context, it has expanded to include the high-value "prompt data" and "completion edits" users generate during interaction. These byproducts are harvested to retrain and "fine-tune" future model iterations via techniques like RLHF (see previous chapter).

the division of labor creates mutual dependence, but bargaining leverage accrues disproportionately to layers where inputs are scarce, difficult to substitute, and embedded in long-lived technical ecosystems. In practice, disentangling these power dynamics is made harder by the opacity of contracts and bundled pricing—cloud capacity, discounts, platform fees, and enterprise support agreements are often negotiated privately—so that the distribution of rents is not transparent even when the underlying dependencies are obvious.

TRENDING TOWARDS VERTICAL INTEGRATION

As in any industry, the build-versus-buy decision shaped enterprise adoption during this period. For example, using a closed-source platform like OpenAI's GPT-4, a firm could access frontier capabilities through APIs and deploy quickly, accepting constraints on architectural modification in exchange for reliability and vendor support (OpenAI 2023). Open model weights offered an alternative path: using releases like Meta's Llama 2, enterprises could modify architectures and tune performance for specific latency and hardware requirements, though at the cost of greater internal engineering effort (Touvron et al. 2023).

While software ecosystems faced these binary choices, the physical hardware layer—particularly in the emerging field of humanoid robotics—mirrored the fractionalized supply chains of the PC era: it was organized around modular "body" components that come together into a complete, functioning machine.[227] At the lowest tier were raw material and subassembly suppliers: precision-machined aluminum and composite shells for limbs and torso frames; high-density PCBs, motors, gearboxes, and power electronics that provide actuation and control. One notch up sits the specialized "hand" segment, where startups such as Proception, Mimic, and

[227] This paragraph draws on Jonas et al. (2025) and Shamsuddoha et al. (2025).

Tacta Systems develop grippers with embedded pressure-and-tactile sensors and multi-axis joint drives. These hands often ship as self-contained modules—complete with onboard sensing, local compute, and firmware—that robot-integrators can bolt onto a standard wrist interface. Parallel to hands are the "arm and torso" integrators—companies like Halcyon Robotics—that take those modular limbs and combine them with mid-body components: shoulder actuators, cable-management harnesses, balance sensors, and real-time control software running on a shared bus. These integrators handle the critical work of kinodynamic coordination (making sure hands, elbows, and shoulders all move in harmony) and package safety interlocks, power distribution, and upstream network interfaces. Above that sat systems integrators and OEMs—Tesla with its Optimus line, 1X Robotics, and others—who sourced hands and torsos from specialist vendors, add their own vision modules (camera arrays, LiDAR), master control software, and battery packs, and then certify the fully assembled humanoid for target applications (warehouse picking, light assembly, consumer service).

However, on balance, the Fourth Industrial Revolution differs from the Third in one critical respect: While the digital/mobile revolution was marked by vertical disintegration—culminating in the decentralized global supply chains analyzed in Chapters 4 and 5—AI's commercialization reversed this trend. By late 2025, the major AI players were moving toward owning more of the stack end-to-end—upstream into data centers, power, and silicon, and downstream into context-aware applications and distribution.[228]

[228] As I will outline below, purely transactional market relationships persisted. However, the broader industry trend favored tighter control. Throughout this section, the term "vertical integration" encompasses both outright internalization (building/owning adjacent layers) and

The smartphone industry demonstrated that modular architectures could enable innovation without integration; the AI industry demonstrated the opposite. When training runs cost hundreds of millions of dollars, when a single supplier controls 90% of critical hardware, when gigawatt-scale power access is severely limited, and when model performance depends on end-to-end system optimization, the logic of vertical integration reasserts itself.[229]

Figure 9.3 maps the AI ecosystem's structure as of late 2025: a network visualization that captures how traditional boundaries between hardware, infrastructure, models, and applications dissolved as companies raced to control multiple layers of the AI stack: chips, browsers, data centers, LLMs, and digital devices. The figure's most striking feature is Nvidia's central position, from which relationships radiate out to virtually every major player. This hub-and-spoke pattern illustrates the industry's fundamental bottleneck: despite hyperscalers' massive investments in internal silicon—Google's TPUs, Amazon's Trainium/Inferentia, Microsoft's Maia, Meta's MTIA, and Apple's NPUs—frontier training and much high-end inference still leaned heavily on Nvidia GPUs manufactured at TSMC.

Figure 9.3 also makes visible the era's defining pattern of "co-opetition" and hedging. Hyperscalers competed at the model and platform layer while simultaneously forming quasi-

---

"quasi-integration"—strategic alliances solidified by long-term exclusivity and equity stakes, such as Microsoft's partnership with OpenAI or Amazon's investment in Anthropic (see AWS 2024c). These deep entanglements replicate the control rights of ownership and differ fundamentally from standard licensing agreements.

[229] For other, similar takes on the AI stack's vertical integration trend see Efrati (2025) and Gardizy (2025).

integrated ties—equity links, long-term compute agreements, and licensing arrangements—to secure scarce capacity and protect against lock-in. Anthropic's dual-cloud positioning with AWS as its primary provider and Google Cloud as a secondary provider, and OpenAI's multi-provider relationships across cloud and silicon partners, illustrate how frontier labs leveraged supplier competition to secure better terms while retaining strategic flexibility.

Finally, the network mapped by the figure highlights how specialized intermediaries occupy the ecosystem's fault lines. Providers such as CoreWeave and Oracle gained leverage by supplying scarce compute capacity when hyperscalers faced constraints, while inference-focused licensing and talent-migration deals signaled a parallel race to reduce the cost of serving models at scale.

The overall picture is an industry moving away from the modular disintegration of the smartphone era and toward vertical reintegration—achieved through equity stakes, long-term supply guarantees, and hybrid partnerships that substituted for the missing markets and incomplete contracts endemic to a capital-intensive frontier. As the following sections detail, most players in the AI stack pursued some degree of integration, transforming the industry from a horizontally specialized ecosystem into a set of competing, vertically interlocking blocs—each spanning from silicon to the end-user experience.

Figure 9.3. Quasi Vertically Integrated AI Stack.

[INSERT FIGURE 9.3 HERE]

Notes: Node sizes reflect relative market capitalization or valuation as of late 2025. Shapes indicate primary function: squares = hardware/silicon suppliers; circles = hyperscalers (hatching indicates in-house chip programs: Google TPUs, Amazon Trainium/Inferentia, Microsoft Maia, Meta MTIA, Apple NPUs); hexagons = cloud/infrastructure providers;

diamonds = foundation model developers; triangles = capital coordinators. Line thickness reflects approximate deal magnitude where disclosed: the OpenAI-Oracle cloud contract ($300 billion) and OpenAI-Nvidia supply arrangements represent the largest commitments, followed by Oracle's $40 billion Nvidia purchase commitment, Microsoft's 27 percent OpenAI stake, Amazon's $8 billion Anthropic investment, and Google's $3 billion-plus Anthropic stake. Relationship types: solid lines = equity investments or chip supply; dashed lines = cloud compute agreements; dotted lines = co-development partnerships; dash-dot lines = strategic partnerships. Some links represent quasi-integration—exclusive capacity contracts, licensing combined with talent migration, or equity-linked supply agreements—that function like vertical integration even without formal ownership. Anthropic's relationships with both Amazon/AWS (primary cloud, $8 billion equity) and Google Cloud (secondary cloud, $3 billion-plus equity) are shown as separate edges to illustrate its dual-cloud hedging strategy.
Sources: Author's analysis based on sources cited in text.

Brief History of Vertical Integration in the AI Stack

Vertical Integration of the AI stack proceeded gradually. Before the AI boom triggered by the introduction of ChatGPT in November of 2022, some larger cloud providers such as Google acquired AI software startups in the wake of accelerated enterprise AI adoption (see Manyika et al. 2017). The AI vertical integration trend really took off during the early 2020s, however, when chip designers began to acquire data-center operators, cloud providers built bespoke AI chips, and hyperscalers incorporated AI software toolchains.

During the late 2010s, before the AI boom really took off, Google expanded into several AI domains, from TPUs to specialized cloud services (see Sato and Young 2017). This precociousness was a downpayment on its emergence as a major player in the AI hardware

space: by the time Google rolled out its seventh-generation TPUs in 2025, they boasted 15–30× better energy efficiency than standard GPUs and allowed tight hardware-model co-optimization tailored for inference workloads (Vahdat 2025). Meanwhile, Google's control over Android—far larger than iOS in most regions—enabled it to deploy Gemini updates directly to billions of devices without app-store review (McCullough 2025). Google also found a way into the end-user, agentic AI game, partnering with Apptronik to embed its Gemini AI into humanoid robots for warehouse and lab use (Apptronik 2024).

Similarly, if not belatedly, OpenAI was particularly aggressive in embedding itself vertically along the AI supply chain. In 2025, its CEO Sam Altman set out on a dealmaking blitz that knit together reciprocal investments, multi-year supply contracts, and crossholdings that tightly tied OpenAI to the hyperscaler and chip ecosystem (Jin 2025). This strategy highlighted the increasingly financialized nature of AI infrastructure. Major deals were struck with Nvidia, involving leasing millions of AI chips, joint data center ventures, and contingency loans. This included a massive five-year, approximately $300 billion cloud services contract with Oracle set to commence in 2027—a deal that spurred Oracle, in turn, to commit $40 billion to Nvidia for the buildout (Welle 2025). OpenAI similarly leveraged its equity stake in the specialized cloud provider CoreWeave to lock in dedicated GPU clusters, bypassing standard allocation queues (Bort 2025b). An agreement with AMD for 6 gigawatts of GPU capacity was similarly financialized, granting OpenAI warrants to purchase up to 160 million of AMD's shares (AMD Inc. 2025). In a further move to diversify its compute supply beyond Microsoft, by mid-2025 OpenAI also added Google Cloud as an infrastructure provider (Jin 2025).

Notably, over the course of 2025, OpenAI became an even more direct participant across the AI stack. A deal with Broadcom was aimed at developing in-house AI chips to cut costs and

reduce dependence on Nvidia's GPU accelerators. The company introduced new productivity features—AI-driven writing suggestions, editing, and document integration—allowing ChatGPT to assist users in composing, formatting, and revising across emails, forms, and documents, a move widely seen as a step toward entering the word processing space (Kahn 2025). Similarly, OpenAI began work on a consumer hardware project, potentially involving an AI-powered device that could serve as a voice assistant or chatbot in the vein of Amazon Alexa or Google Home (Williams 2025).

By the end of 2025, OpenAI had also expanded into a high-end consulting model by embedding its own engineers onsite with clients.[230] On the one hand, the company fine-tuned and deployed custom AI solutions, adapting OpenAI models to proprietary data and workflows. On the other hand, it helped build end-user applications such as specialized chatbots, supervising data-labeling by domain experts in fields like medicine and integrating AI into existing IT stacks. The target clients for these services were major enterprises and government agencies, including a reportedly $200 million contract with the DoD for planning and operational AI tools (Holmes and Muppidi 2025), custom GPT-4o Vision fine-tuning for Grab to enable automated map-labeling using hundreds of street-level images (ibid), and a bespoke chatbot for Morgan Stanley to surface internal research (Son 2025).

Meanwhile, during this same time Amazon had also consolidated control over several stages of the AI stack. AWS doubled down on its custom AI silicon with Trainium for training and Inferentia for inference to wrest some GPU spend back from Nvidia and offer lower-cost, higher-throughput AI applications. AWS also launched its own large language models such as

[230] This passage draws extensively on Holmes and Muppidi (2025).

AWS Titan so it need not rely solely on OpenAI or Anthropic for AI-as-a-service. In 2025, Amazon even blocked third-party AI agents—such as Google's Project Mariner—from crawling or transacting on its site, prioritizing its own assistant, Rufus, to monopolize the entire process by which consumers discover and buy products (Gehan 2025).

In the early 2020s, Meta moved to vertically integrate its AI stack—building custom silicon and full-stack infrastructure to align data, model training, and deployment (Meta 2024). By licensing its open-source Llama models through third-party cloud platforms like AWS and Azure, and introducing the Llama API service at LlamaCon 2025, Meta cultivated broad deployment and developer adoption (Cai 2025). Meta also accelerated toward proprietary AI search capabilities—developing an in-house crawler and real-time data engine aimed at powering its AI assistant independently of Google and Microsoft. Finally, the company took a 49% stake in Scale AI to secure greater control over training data and labeling infrastructure (see ScaleAI 2025).

Nvidia followed a similar playbook.[231] In mid-2025, the company executed this pivot by acquiring Lepton AI and launching DGX Cloud Lepton. This new "marketplace" aggregated capacity from Nvidia's own fleet and over a dozen specialized partners—including CoreWeave and Nebius—to offer developers "instant access" to both GPUs and bundled software tools. While initially focused on the heavy lifting of model training, Nvidia positioned the platform to compete on inference—running finished models more cheaply—posing a direct challenge to

[231] This paragraph draws on Janakiram (2025); Nvidia (2025b); Seetharaman and Sriram (2025).

Amazon's specialized chips (Nvidia 2025c). Finally, to further cement its position in the stack, the company made a sizable investment in OpenAI.[232]

Nvidia also sought licensing deals and executive migration as a distinct vertical-integration like mechanism. In late 2025, it struck a nonexclusive licensing deal with inference-chip startup Groq, alongside an unusual organizational move: Groq's founder-CEO Jonathan Ross and other senior personnel joined Nvidia to help integrate the licensed inference technology, while Groq continued operating independently (see Clark 2025). Thus, the AI behemoth moved to internalize a key complement to its training-dominant GPU franchise without a full acquisition, effectively importing both specialized IP and human capital into its own product stack.

Microsoft pursued vertical integration as well, though with a strategy distinct from the "total control" of Google or Apple. Microsoft converted its prior economic rights in OpenAI into a conventional equity stake of roughly 27% in the newly recapitalized OpenAI Group Public Benefit Corporation, while extending key partnership terms through 2032 (Microsoft 2025a). Microsoft also aggressively deployed its own silicon—the Azure Maia series.

SoftBank offers a different—but revealing—path toward vertical integration: coordination through capital rather than through a single operating firm. Unlike Google or Apple, which integrated capabilities within corporate boundaries, SoftBank assembled a portfolio spanning critical layers of the AI stack while keeping portfolio companies operationally independent (see Jie 2025). It owns Arm Holdings, whose low-power architecture underpins

[232] While Nvidia initially floated making a massive $100 billion infrastructure megadeal with OpenAI, by early 2026 the AI chip designer proposed a more targeted equity-led partnership.

edge AI and hyperscale processors, while simultaneously moving to develop its own AI-specific semiconductors. It made large strategic investments in OpenAI and partnered with system-level vendors like SambaNova. This approach, using ownership and capital allocation to manage technological interdependence where markets and contracts remain thin, represents an alternative organizational form for navigating the AI stack's complexity (SoftBank Group Corp. 2025; Business Wire 2025).

Apple's Hybrid Strategy

Apple also pursued AI self-sufficiency to a certain extent, but its strategy was a hybrid approach that combined its traditional emphasis on on-device intelligence with a new, pragmatic move to the cloud. Moreover, during the early 2020s Apple was also somewhat of an outlier: rather than acquiring AI capabilities it lacked or building them from scratch, it often went out into the open market and, when needed, simply forged old-fashioned deals across the supply chain.

To be sure, the company long emphasized on-device intelligence to prioritize user privacy, integrating hardware, software, and edge computing to process data locally—exemplified by its development of the Neural Processing Unit (NPU), the chip inside every iPhone that handles tasks like real-time translation, and its 2020 acquisition of on-device AI startup Xnor.ai (Boyle et al. 2020). However, by 2024-2025, Apple was preparing to power its new "Apple Intelligence" cloud services through a dual-pronged infrastructure strategy.

First, to handle future training and complex inference workloads at scale, Apple began co-designing a custom server-class AI chip with Broadcom (Ma and Liu 2024). This move complemented its edge strategy rather than replacing it, with future iPhones slated to delegate simple on-device tasks to their NPU while seamlessly handing off heavier workloads to this new server silicon.

Second, for immediate compute-intensive tasks, Apple deployed Private Cloud Compute. This proprietary server architecture builds on the company's end-to-end encrypted iCloud foundation to verify that user data is deleted immediately after processing, ensuring that requests are not stored or accessible even to Apple staff (Apple 2024c). Simultaneously, the company quietly explored turning its in-house server infrastructure into a broader resource. Reports indicated Apple was testing clusters of its own M2 Ultra chips in data centers—essentially Mac-mini–style servers (Gurman 2024).

Ironically, however, despite its reputation for end-to-end control (see Chapter 6), Apple was arguably the furthest from controlling the AI stack among major platform companies in 2025. While Apple dominated the distribution layer through iOS defaults and the App Store, and its custom silicon provided powerful on-device capabilities, it remained comparatively weak at the foundation model and data retrieval layers—lacking both a web-scale index like Google's and a frontier model lab (see Tilley 2024). Apple instead struck a non-monetary "Apple Intelligence" deal to plug ChatGPT into Siri rather than constructing a homegrown LLM for that specific feature, essentially outsourcing the 'brain' of the operation while retaining control of the interface (Apple Inc. 2024b).

Putting it All Together

Table 9.1 synthesizes these varied approaches. While all major players moved toward greater vertical integration during the early 2020s, they pursued distinct strategies shaped by their existing assets, competitive positions, and organizational capabilities. Google leveraged decades of infrastructure investment to build organically across nearly every layer. Microsoft and OpenAI relied on equity relationships and contractual arrangements to assemble capabilities without full operational control. SoftBank coordinated through capital allocation rather than

integration within corporate boundaries. Apple, despite its reputation for end-to-end control in consumer electronics, remained the most dependent on external partners at the model layer. These differences mattered: they shaped not only cost structures and margins but also the pace of innovation, the distribution of bargaining power, and the durability of competitive advantage across the stack.

Table 9.1. Strategic Approaches to AI Stack Integration

Firm	Integration Strategy	Primary Mechanism	Key Moves	Relative Dependencies
Google	Total control	Organic build	TPUs, Gemini, Android distribution, Apptronik robotics partnership	—
Amazon	Full-stack via AWS	Build + investment	Trainium/Inferentia, Titan models, $8B Anthropic stake, Rufus agent	Frontier model leadership
Microsoft	Platform + partnership	Equity-anchored	27% OpenAI stake, Azure Maia, GitHub Copilot	Proprietary frontier model

Firm	Integration Strategy	Primary Mechanism	Key Moves	Relative Dependencies
Meta	Open-source leverage	Organic + strategic stakes	MTIA chips, Llama licensing, 49% Scale AI stake, in-house search	Cloud infrastructure
Apple	Hybrid edge-cloud	Selective partnerships	NPUs, Private Cloud Compute, ChatGPT-Siri deal	Foundation models, web-scale data
Nvidia	Downstream expansion	Licensing + services	DGX Cloud Lepton, Groq talent/licensing deal, OpenAI investment	Application layer
OpenAI	Aggressive embedding	Deals + partnerships	$300B Oracle contract, Broadcom chip deal, Atlas browser, CoreWeave equity	Proprietary silicon
SoftBank	Coordination through capital	Capital allocation	Arm ownership, OpenAI investment, SambaNova partnership	Operational integration

Notes: Table summarizes dominant integration strategies among major AI stack participants as of late 2025. "Relative Dependencies" column reflects author's analysis of layers

where each firm remained reliant on external suppliers or partners based on capabilities described in text.

Sources: Author's analysis based on sources cited in text.

Making Sense of the AI Stack's Evolving Structure

If foundation models are the scarce, capital-intensive input in the AI supply chain, why do the labs that build them also operate consumer applications rather than licensing exclusively through APIs—and why have the lines blurred between AI hardware and software? The evolution toward the byzantine market structure depicted in Figure 9.2 makes considerable economic sense. In AI, vertical integration is less a managerial fad than a rational response to three forces that operate simultaneously: double marginalization dynamics, contracting frictions, and cross-layer complementarities. Conversely, where those conditions do not hold, as in the example of humanoid robotics circa 2025 that I will outline further below in this section, vertical disintegration and modular specialization persist.

Double Marginalization Dynamics

First, cost discipline at massive scale helps explain vertical integration in the AI stack. While standard economic theory suggests specialization optimizes scale, the unique market structure of AI hardware inverts this logic. In a competitive market, buying from a specialized supplier is efficient.[233] But the AI stack features market power at multiple layers. The dominant chip

[233] Standard economic theory posits that vertical disintegration lowers industry-wide costs through the aggregation of demand. In a disintegrated market, a single specialized supplier (e.g., a pure-play foundry like TSMC) combines orders from N downstream competitors. This allows the supplier to reach a total production volume ($Q_{\text{total}} = \Sigma q_{\text{i}}$) that far exceeds the volume

supplier commands gross margins exceeding 90 percent (Jon Peddie Research 2024). In terms of cloud providers, while they compete fiercely with one another, they also exercise pricing power by bundling scarce GPU capacity with storage, governance, and workflow integration—and by using forecasting and orchestration to allocate reliable access during demand surges (Katsurashima et al. 2024).

Consequently, when a dominant accelerator supplier and a dominant infrastructure layer each impose their own markup, the vertical chain suffers from successive markups (double marginalization): independent pricing decisions raise the total price and restrict output relative to what would be chosen by an integrated firm maximizing joint profit (Spengler 1950).[234] Vertical integration—or equivalent contractual remedies—can reduce this inefficiency by internalizing pricing decisions across the vertical chain and eliminating successive markups (Spengler 1950).

Therefore, perhaps the most persuasive empirical support for the double-marginalization hypothesis lies in the simultaneous rush by every major hyperscaler to develop proprietary

---

achievable by any single integrated firm (q_{i}). By spreading appreciable fixed costs (F) over this larger quantity, the specialized supplier pushes production further down the falling portion of the long-run average cost curve (LRAC), achieving a lower unit cost than any individual firm could replicate: $AC(Q_{\mathrm{total}}) < AC(q_{\mathrm{i}})$.

[234] In a disintegrated supply chain, the upstream firm (e.g., chip designer) and downstream firm (e.g., cloud provider) each price strictly to maximize their own slice of the profit at Lerner Margins (see Chapter 4) and ignoring the harm their markup causes to the final demand. This results in a "bad" Nash Equilibrium: a final price that is too high (Price > Monopoly Level) and a total quantity that is too low.

silicon—Google's TPU, Amazon's Trainium/Inferentia, Microsoft's Maia, and Meta's MTIA. While typically marketed as a technical optimization (see, e.g., Borkar et al. 2024), the economic driver is clear: by internalizing the chip layer, these firms seek to bypass the extraordinary margins earned by Nvidia, thereby reducing the upstream markup embedded in their cost structure. Reciprocally, Nvidia sought to capture a greater share of the value chain by launching DGX Cloud Lepton to rent compute directly to enterprises. By moving beyond its historical role as a mere supplier to intermediaries like Dell or AWS, Nvidia effectively bypassed the cloud providers' own markups to capture downstream margin for itself (see Nvidia 2025c). In both directions, the boundary of the firm shifted toward swallowing the adjacent layer to eliminate the "double tax."

The definitive signature of this integration is the dramatic expansion of capacity—the remedy for the supply rationing predicted by Spengler (1950). The massive capital expenditures on projects like the OpenAI-led "Stargate" initiative, which I will elaborate upon further below, suggest that quasi-integrated firms are choosing to deploy compute volume far beyond what an arm's-length market would support. Likewise, once chips, data centers, and models are treated as a single cost-optimized unit, internal cost per inference token and training FLOP falls enough to justify buildouts that would look irrational if each layer were purchased at full retail markup (Affinius Capital 2025).

Transaction Costs and Bargaining Frictions

Second, vertical integration acts as the ultimate insurance policy against "holdup" (Williamson 1985). In the AI stack, components are rarely plug-and-play; data pipelines, model architectures, and inference chips are often deeply customized to work together. "Co-specific assets" of this ilk carry a massive risk: if a firm spends billions tuning its models for a partner's specific hardware,

it becomes trapped. The partner, knowing the firm cannot easily switch, can suddenly demand higher prices or unfavorable terms (Klein et al. 1978). By owning the entire stack, a company eliminates this vulnerability. The same logic applies to enterprise customers: once a business rewires its daily workflows around a specific AI platform, switching becomes a technical nightmare. AI giants therefore faced a stark choice: remain dependent on potential rivals for critical inputs, or invest hundreds of billions to control their own destiny (Döpper et al. 2024).

Consider the strategic logic behind Microsoft's strategic partnership with CoreWeave. While Maia absorbed a growing share of Microsoft's internal inference workloads for offerings like Copilot, the company still relied heavily on Nvidia clusters for training frontier models. For Microsoft, forging a long-term supply agreement with CoreWeave eased capacity constraints while Azure underwent retrofitting. And for its strategic partner, OpenAI, securing a large, multi-year commitment from a specialized GPU cloud provider diversified compute access beyond dependence on a single hyperscaler for training and running models (Hammond et al. 2025).

Then there is the fact that user interactions generate the preference data and usage logs that flow back upstream to improve models. Distribution is therefore not merely a revenue channel but a strategic asset for data acquisition—a dynamic that helps explain why OpenAI, Anthropic, and Google all maintain direct-to-consumer interfaces alongside their API businesses. By owning a browser, foundation model developers like OpenAI ensure that ChatGPT users don't have to rely on Google Chrome, preventing their primary rival from throttling access or observing user data (McMahon 2025).

Similar examples abounded during this era. Microsoft's acquisition of GitHub positioned it to bundle Copilot with the platform where millions of developers already worked, creating

distribution leverage that independent competitors struggled to match. Meanwhile, Cursor's rapid growth attracted acquisition interest from OpenAI and xAI—not primarily for its revenue, but because its users generated precisely the kind of expert coding data (reasoning traces) that labs needed for model improvement (Dzieza and Field 2025). Similarly, Anthropic pursued a dual-cloud strategy, securing $8 billion in total investment from Amazon (making AWS its primary cloud provider) while also raising over $3 billion from Google and maintaining Google Cloud as a secondary provider (Amazon Staff 2024; Field 2025a).[235]

Finally, Nvidia's "license-plus-talent" deal with Groq mirrors a broader pattern in the AI ecosystem in which incumbents capture capabilities and reduce dependency through partial integration. As inference demand rose sharply—and energy efficiency became a binding constraint—Nvidia pounced with its acquihire-partnership strategy, underscoring the fact that, when speed and control matter, yet full mergers are costly, slow, or politically constrained, there are creative ways of approximating full vertical integration.

Coordination Economies

Third, coordination economies also help explain vertical integration in the AI stack. R&D spillovers increasingly run across the value chain. Performance improvements depend on joint optimization of hardware, software, and deployment—so integrating adjacent layers accelerates iteration and improves quality (see Langlois and Robertson 1995). In the limit, integration is not

[235] Another explanation for vertical integration in some cases was plain-old vanilla upstream capacity constraints: Frustrated by slow capacity turnarounds from Oracle Cloud and other providers, xAI built the Colossus data center rather than rely exclusively on hyperscalers for training its frontier models.

just about lowering costs; it is about building full-stack systems and devices capable of persistent, context-aware inference—where sensors, edge compute, cloud capacity, and model updates form a single integrated product.

Take some representative examples. Microsoft paired its Maia AI chips with a custom rack design and a dedicated "Sidekick" liquid-cooling unit to increase rack density and energy efficiency in AI data centers (Borkar et al. 2024). For their part, Oracle, AMD, and SoftBank launched multi-billion-dollar co-investments to codevelop AI-optimized chips (Forgash and Ghosh 2025).

Choosing Not to Vertically Integrate Provides Further Insights

As a point of contrast, take the Humanoid Robot supply chain I outlined above. Its fragmented structure is not accidental; it is precisely what one would expect in the absence of the forces driving vertical integration in the AI stack. Because key components remain modular, assets are largely redeployable, and no single supplier exercises dominant pricing power across adjacent layers, neither double marginalization nor holdup risks loom large. Innovation can therefore proceed through specialization and arm's-length contracting rather than internalization. The contrast with AI is instructive: where model performance depends on tightly co-optimized data pipelines, hardware, and deployment infrastructure—and where market power is exercised at multiple points—firms are pushed toward vertical integration.[236]

Unsurprisingly, therefore, over the early 2020s, each new AI model generation demanded larger clusters of GPUs, more advanced networking, and vastly higher energy budgets, which

[236] Of course, should humanoid robotics come to depend on tightly coupled, large-scale AI training and inference infrastructure, similar integration pressures may yet emerge.

blurred the line between R&D and CaPEx, transforming laboratory science into a form of heavy industry. The next chapter details the resulting AI CapEx wave—tracking the enormous spending on data centers, accelerators, networking, and power infrastructure that made the AI boom possible.

CONCLUSION

This chapter argues that understanding AI requires understanding its industrial organization—the supply chain structure, bargaining dynamics, and capital requirements that determine who can compete at the frontier and who captures value. The vertical integration trend documented here reflects a structural shift in how the technology industry organizes itself, with implications extending far beyond AI narrowly defined. Today's AI boom is not simply the emergence of a powerful new technology, but the rapid construction of a new industrial stack—one whose economics are reshaping firm boundaries, business models, and supply chain bargaining power. Like electricity and semiconductors before it, AI exhibits the exponential improvement curves characteristic of general-purpose technologies (see Chapter 2)—but what distinguishes AI in 2025 is the sheer scale, cost, and sophistication of its physical and digital moving parts.

The modern deep-learning paradigm initially rewarded scale: many early leaps in performance were purchased by orders-of-magnitude increases in compute, ever-larger training runs, and capital-intensive infrastructure. At the same time, the critical inputs that make scaling possible—frontier chips, data-center capacity, electricity, high-end networking, scarce talent, and proprietary data pipelines—became binding constraints rather than background conditions.

These constraints pushed the industry toward vertical integration and tighter ecosystem control. When key bottlenecks sit upstream (specialized GPUs, cloud capacity, model-training infrastructure) and downstream (distribution, proprietary data, application-layer lock-in), firms

face strong incentives to internalize more of the stack, build closed interfaces where it suits them, and use contractual and technical governance to coordinate dependencies across hardware, cloud infrastructure, models, and applications. In other words, circa 2025, the AI era revived an old industrial lesson: where fixed costs soar and bottlenecks harden, organizational form becomes strategy—and the boundaries of the firm shift accordingly.

Moreover, the AI frontier is no longer defined by scale alone. As the previous chapter outlined, as marginal returns to simply adding parameters, tokens, and FLOPs diminished, leading firms increasingly sought gains through compute-efficient innovation: improved training and post-training methods, synthetic and curated data strategies, retrieval-augmented systems, multimodal architectures, tool-using and agentic workflows, and specialized models optimized for particular tasks and deployment environments.

But AI's trajectory will be shaped as much by technical advances as by organizations learning how to adapt to it. While the productivity potential of AI is real, it ultimately depends on diffusion and commercialization: whether interoperability emerges to reduce redundancy and switching costs; whether systems solve real enterprise and consumer problems at an affordable price in time and money; and whether firms learn to redeploy human attention away from tedious, automatable tasks and toward creativity, judgment, and strategy.

Therefore, Chapter 11 turns from the anatomy of the AI stack to its consequences. I examine how a wave of AI applications may bolster economywide productivity, reallocate value, and transform job markets—and, drawing on earlier industrial revolutions, how AI may reshape industries, create new occupations, and displace some workers. But before we get to that point, it's important to revisit the fundamentals of industrial revolutions and the more generic type of GPTs—general purpose technologies. That is what the next chapter does.

CHAPTER 10. AI STANDARDIZATION & ORGANIZATIONAL ADAPTATION

If the early 2020s were defined by the promise of generative AI—dazzling demos, breathless hype, and frantic experimentation—the middle of the decade was defined by its industrialization. Things moved from a period of "wildcatting," characterized by scattered pilots and bespoke, fragile implementations, to a phase of massive capital commitment and structural hardening. By 2025, the sheer magnitude of the physical build-out—hundreds of billions of dollars poured into data centers, energy grids, and silicon foundries—signaled that the global economy was no longer merely flirting with artificial intelligence; it was pouring the concrete foundation for a new techno-economic paradigm.

This shift was driven by a powerful economic logic: the transition from scarcity to abundance. As the cost of "intelligence" (inference) plummeted and the friction of deployment declined materially through nascent standardization efforts, AI began to behave less like a luxury good and more like a utility. Yet, this transition was not instantaneous. It required a massive, coordinated retrofit of the industrial base—everything from the mines for copper and fabs for chips to the factories for cooling systems and the assembly of servers—as well as a re-wiring of energy infrastructure, and, perhaps most difficult of all, a fundamental reorganization of the modern enterprise itself. It was now incumbent upon firms like General Motors, JPMorgan, and Walmart to find ways to harness the potential of AI to improve their processes and offerings to better compete and satisfy their customers' evolving demands.

The reorganization efforts began with a palpable shock to the United States' physical stock. The release of ChatGPT by OpenAI in November of 2022, an easy-to-use, mass-market conversational interface to large language models, set off an explosive surge in construction

spending for the core infrastructure behind artificial intelligence.[237] In less than three years, an extraordinary wave of new plants making specialized AI chips and electronics got built across the US, along with new data centers to house that hardware; a parallel expansion of the electric grid to power it all also occurred.

Consider Figure 10.1, which graphs this phenomenon. The "Computers, Electronics & Electrical" sector really stands out: while it was once a relatively small portion of industrial construction, between 2023 and mid-2025, when the data series ends, it is the single largest recipient of investment in the construction of new structures. Composed of specialized facilities that make semiconductor chips, routers, and network switches, this category grew by 52% after 2022, adding nearly $40 billion in new annual spending. Importantly, these facilities include those busy making, testing, or packaging GPUs.

The figure also shows that this boom was matched with an even more explosive, 151% surge in "Data Center" construction, which provided the secure, climate-controlled buildings necessary to house all the new hardware that creates and runs AI. Think of a data center as a high-tech factory for the digital age. Instead of manufacturing physical goods, its purpose is to process, store, and distribute the immense amounts of information that power our online world. Inside these highly secure, climate-controlled buildings are tens of thousands of interconnected devices stacked in racks from floor to ceiling: powerful computers called servers that perform calculations, vast arrays of storage devices that hold the data, and high-speed network equipment

[237] While OpenAI's image-generation systems (e.g., DALL·E) predated ChatGPT, the chatbot made interactive text generation mainstream at unprecedented scale, and later upgrades broadened its multimodal capabilities.

that directs the flow of information. Everything is supported by a massive life-support system: redundant power grids to prevent blackouts, high-speed fiber optic networks for connectivity, and advanced cooling systems to keep the hardware from overheating (Green et al. 2025).

Figure 10.1 The Epic AI Infrastructure Construction Boom.

[INSERT FIGURE 10.1]

Notes: The figure depicts several categories for Value of Private Construction Put in Place (VIP), which measures the monthly dollar value of construction work installed on site, encompassing materials, on-site labor, a share of equipment rental, architectural and engineering work, and other project costs, while excluding maintenance and repairs, land costs, and production machinery or equipment such as servers, racks, or GPUs. These data are reported as seasonally adjusted annual rates (SAAR) in millions of dollars, converted here to billions. Data center construction falls under the office category and focuses on structures for computing operations, including the shell along with power, cooling, and integrated building systems for facilities that store, process, and transmit digital information. The data source (disclosed below) began publishing a separate data center line in 2024 with backdated history to January 2014, prior to which such spending was embedded within general office construction, creating a classification quirk where pre-2014 office spending baselines are not directly comparable to post-2014 figures. The computers, electronics, and electrical category corresponds to manufacturing plant structures under NAICS 334 (computers, semiconductors, communications gear) and NAICS 335 (electrical equipment and appliances), representing factory buildings and improvements but not production tools. Manufacturing spending, excluding NAICS 334 and 335, includes all other manufacturing structures such as chemicals, transportation equipment, fabricated metals, plastics and rubber, and nonmetallic minerals, with

the source classifying all buildings at manufacturing sites including on-site warehouses under this category. The electric overlay is drawn from the power category and includes generation plants, substations, switch houses, transformers, and transmission lines relevant to data center clusters, omitting gas and oil elements. VIP data components are computed separately before stacking, including manufacturing (total minus NAICS 334 and 335), computers, electronics, and electrical (NAICS 334 plus 335), data centers, and electric as an overlay not included in the stack. Each monthly series is smoothed using a three-month trailing simple average applied after conversion to real dollars, with nominal SAAR series deflated to constant June 2025 dollars using the seasonally adjusted CPI-U (for urban consumers) via the standard ratio method. VIP data undergo seasonal adjustment using concurrent factors that are revised as new data arrive along with periodic benchmark revisions. Coverage is limited to private VIP only and includes new construction and major improvements with on-site mechanical and electrical installations like HVAC and elevators.

Sources: U.S. Census Bureau 2024a; U.S. Census Bureau, 2024b; U.S. Census Bureau, 2024c; Federal Reserve Bank of St. Louis 2025.

While data centers powered the internet for decades, the rise of generative AI placed them at the epicenter of a new industrial boom. Training and running large AI models requires an astonishing amount of compute—the combined hardware, processors, and energy needed for intense calculations. This workload creates extreme power density, a massive concentration of energy consumption within each server rack. To handle this load, a data center's electrical system must be upgraded with components like larger switchgear and floor-mounted power distribution units to safely deliver the increased wattage. This intense density also causes the hardware to get physically hot, forcing the industry to shift from traditional air conditioning to more effective

liquid cooling systems, which circulate fluid directly over the processors to draw away heat, much like a car's radiator (see Donnellan et al. 2024).

To support this ecosystem, the “electric grid” saw a parallel 15% build-out. In contrast, construction in the broader “Manufacturing” category grew by a modest 14% during this period.[238]

What’s remarkable is that Figure 10.1 does not cover spending on equipment like GPUs themselves. Nor does it identify spending on AI related physical structures outside of the US, even though most AI semiconductors are fabricated in Taiwanese chip factories. It merely shows construction in the plants and similar physical infrastructure associated with three AI adjacent categories. Yet, the figure nonetheless echoes earlier industrial revolutions: it documents how a “heavy-infrastructure-first” footprint preceded the mass diffusion of technologies that forever changed the world.

Yet, infrastructure is merely the prelude to impact. Just as the significance of the railroad lay not in the tracks laid down but in the new markets they connected, and just as the internet’s story was not the unfurling of endless fiberoptic cable, the true economic story of the mid-2020s was not the construction of data centers, but the rapid diffusion of the intelligence they generated into the fabric of the real economy.

---

[238] While the figure’s "Manufacturing" category almost certainly contains "AI-adjacent" suppliers—for instance, the plants that build the essential cooling systems, machinery, and metal racks for data centers—it is dominated by transportation equipment (likely linked to Electric Vehicles) and chemicals, which together accounted for over $9 billion in new construction spending after 2022.

In this vein, this chapter also assesses several important AI application areas—healthcare, manufacturing, finance, and education and research—and examines how as AI standardizes and becomes more plug-and-play, it may unlock significant economic transformations. Near-full automation in factories and warehouses approaches viability, shifting human labor to higher-value supervisory roles, while supply chains move toward frictionless, dynamic optimization. In healthcare and biotech, standard AI solutions for diagnostics and drug discovery are compressing development timelines, leading to faster breakthroughs and personalized medicine. Meanwhile, automated financial services are reducing overhead to expand inclusion, and AI-driven research tools are beginning to accelerate the very rate of scientific discovery itself.

This distinction matters greatly. We know from previous industrial revolutions that the commercialization of General Purpose Technologies redefines not only patterns of labor and leisure, production and consumption, but also how societies undertake discovery, generate knowledge, and conceive of progress (Mokyr 1990; Gordon 2016). If the current trajectory holds, the Fourth Industrial Revolution promises to supercharge these transformations in ways that far exceed its predecessors. For that reason, the next chapter turns to this expanded horizon, exploring the potential long-run economic impacts of AI's standardization and widespread adoption informed by the history of foundational technologies.

The rest of this chapter proceeds as follows. First, I quantify the historic capital injection into the AI ecosystem, documenting the "hockey-stick" surge in capital expenditure by hyperscalers and startups alike, and the physical constraints—from energy to cooling—that this build-out encountered. Second, I analyze the mechanisms that drove early AI diffusion, specifically the plummeting marginal cost of inference and the emergence of interoperability standards (from ONNX to the Agentic AI Foundation) that turned fragmented tools into a

cohesive, if politically contested, supply chain. Finally, I examine the "ground truth" of adoption, moving beyond the hype to assess how specific sectors—manufacturing, healthcare, and finance—were navigating the complex interplay between technical standardization and the organizational reforms required to unlock AI's potential during 2025. I also speculate how continued progress towards AI integration by these industries may look like into the near future.

HOW MUCH HAS BEEN INVESTED, IN WHAT, AND BY WHOM?

By late 2025, the AI boom had begun to register some direct economic effects. Consider some representative examples. OpenAI's revenue run rate reached roughly $12 billion in 2025, and ChatGPT had on the order of 700 million weekly active users; Anthropic's large enterprise accounts—each representing more than $100,000 in annual revenue—had grown sevenfold year-over-year (PM Insights 2025).[239] While xAI also surged, leveraging its integration into the X platform to rapidly scale subscription and ad-linked revenues (Sacra 2025a), regulated industries adopted Cohere's privacy-centric models at a steady clip, propelling its annualized recurring revenue to approximately $150 million (Sacra 2025b).

Yet, looking at these firms through the lens of direct employment reveals a stark contrast to previous industrial giants. By late 2025, the combined headcount of the sector's leaders remained surprisingly lean: public estimates placed OpenAI in the low thousands (Unify 2025a), Anthropic and xAI near 1,000 to 1,200 each (Austin 2025; Kay 2025), and Cohere at roughly 800 (Unify 2025b). This generated revenue-per-employee figures that dwarfed those recorded

[239] Revenue run rate is a financial metric that projects a company's future annual revenue by extrapolating its current revenue over a shorter period, assuming the same performance continues.

during the so-called SaaS era, signaling a decoupling of value creation from direct labor headcount at the technological frontier.[240] This hyper-efficiency creates a paradox that will be explored in Chapter 11: while the AI production layer requires remarkably few workers, the AI deployment layer promises to reshape the labor market for millions. For now, however, the immediate economic story was not about labor, but capital—which registered important, economy wide effects.

While the software layer remained lean, the infrastructure layer required to support it was ballooning into a capital-intensive behemoth with palpable macroeconomic spillovers. AI and data center-related CapEx reportedly approached $400 billion in 2025, stoking a build-out that represented the largest private-sector capital wave since the tech cycle of the late 1990s (see Bratton 2025; Mims 2025b).[241] Indeed, by November of that year, AI-related outlays were on

[240] The "SaaS era"—roughly 2005–2022—refers to companies like Salesforce, Slack, and Workday selling cloud-hosted software subscriptions rather than physical licenses, representing the pinnacle of high-margin, scalable tech. Even best-in-class SaaS firms typically generated between $300,000 and $600,000 in revenue per employee, reflecting the need to scale human infrastructure across sales, support, and engineering (Kehayias 2023). The frontier AI firms of 2025, by contrast, reportedly generated figures exceeding $5 million to $10 million per employee—illustrating a shift where revenue growth is constrained by capital (e.g., chip availability) rather than labor (Cuofano 2025).

[241] The widely cited "nearly $400 billion" figure refers to nominal, calendar-year CapEx projections (including finance leases) for Alphabet, Microsoft, Amazon, and Meta and reflects a broader definition of capex than "AI-only" infrastructure.

pace to add at least 0.5 percentage points to U.S. real GDP growth for the year (see Lascelles and Nye 2025). Incredibly, investment in information-processing equipment and software—the category capturing the servers and gear driving this surge—accounted for roughly 92% of U.S. GDP growth in the first half of 2025 (Furman 2025).

To put this magnitude into perspective: these investment numbers rivaled or even exceeded the growth contribution of consumer spending in the second quarter, despite the latter remaining the dominant component of the absolute economy. Projections held that this momentum would persist, contributing up to 0.5 percentage points to growth through 2026 (Lascelles and Nye 2025; Winkler et al. 2025). Not since the peak of the dot-com boom had a business investment category outside of consumption exerted such hydraulic pressure on economic momentum; in fact, U.S. tech capital expenditure grew more in the five years ending in 2025 than during the entire 1999–2002 bubble (see Rana and Girard 2024).

What's more, by year's end, the "CapEx blowout" appeared to have further room to run. Driven by a physical bottleneck that reshaped competitive dynamics, it presaged a massive retrofit cycle that promised to sustain spending well into the late 2020s (Maslej et al. 2025). By 2025, most extant data centers could not supply enough power for the newest AI chips: while legacy racks averaged under 8 kW, frontier AI systems demanded over 30 kW (Donnellan et al. 2024). While startups like CoreWeave and xAI could build purpose-designed facilities from scratch, incumbent hyperscalers faced the slower work of upgrading existing infrastructure.

The Brief History of the AI Boom

Even before the AI investment frenzy of the early 2020s, Big Tech companies, venture capital funds, and governments invested heavily in AI. As early as 2016, Google, Baidu, and other major tech companies invested an estimated $20–30 billion in AI R&D (Columbus 2017). Additionally,

AI startup funding reached $15.2 billion in 2017, with China receiving nearly half of this funding (48%) and the US capturing approximately 38% (CB Insights 2018)—a pattern that was subsequently reversed with the advent of generative AI.

Global corporate investment in AI—spanning mergers and acquisitions, minority stakes, private investment, and public offerings—rose sharply during the early 2020s (Maslej et al. 2024; Maslej et al. 2025). Figure 10.2 documents hyperscaler capital expenditures between 2019 and 2025 in real 2024 dollars. Combined CapEx by AWS, Google, Microsoft, and Meta rose steadily from about $86B in 2019 to $161B in 2022, dipped slightly to roughly $144B in 2023—as the sector optimized for efficiency in the wake of the COVID-19 hiring frenzy that overtook the tech sector—and then surged to $218B in 2024. This produces a clear "hockey-stick" profile: a gradual pre-2022 climb, followed by a sharp re-acceleration after the generative AI inflection. This shift is consistent with an infrastructure arms race centered on data centers, accelerators, power, cooling, and networking rather than incremental software-only scaling.

The pivot is best captured by the revenue of Nvidia's data center business—the industry's primary proxy for infrastructure demand—which tripled in a single fiscal year, skyrocketing from roughly $15 billion in fiscal 2023 to $47.5 billion in fiscal 2024 (Nvidia Corp. 2024). This marked the moment the industry shifted from purchasing relatively modest compute for research labs to procuring it for industrial-scale factories in ample quantities.

Figure 10.2 Hyperscaler CapEx Spending, 2019-2025.

[INSERT FIGURE 10.2]

Notes: The figure plots annual, combined capital expenditures (CapEx) for AWS, Google, Microsoft, and Meta. CapEx figures are from company disclosures and include finance leases; 2025 totals are estimates. The lightly shaded region marks the post-2022 period in which

hyperscaler CapEx re-accelerated. Values are shown in real 2024 dollars, obtained by deflating nominal CapEx using the CPI-U (All Items, U.S. city average; 1982–84 = 100). For 2025 deflation, the 2025 CPI-U annual average are used, but are provisional: they are computed as the simple average of the available monthly CPI-U index levels from January–September 2025. Sources: Affinius Capital 2025; BLS 2025a.

As this monumental transition took hold, money poured in to defray exorbitant training costs for generative AI models, such as the estimated $100 million for OpenAI's GPT-4 and close to $200 million for Google's Gemini-Ultra (Maslej et al. 2024: 63). It also helped shoulder the increased computational demands associated with AI inference. Consider that, circa 2025, a single typical "inference" (that generates a response to a user query) can consume 3×10^{11} FLOPs.[242] At a cost of $0.002 per 1,000 tokens (roughly 750 words), and assuming an average response length of 100 tokens, a service handling 1 million user queries per day incurs inference costs of approximately $200/day. For larger services handling billions of queries daily, this cost runs into the tens of millions of dollars.

2025: The Spending Dam Bursts

Combined, CapEx spending by the "big four", Microsoft, Google, Amazon, and Meta, exceeded $320 billion in 2025, with nearly $100 billion spent in the second quarter alone (Bratton 2025; Winkler et al. 2025)—rising from approximately $151 billion in 2024 (Lascelles and Nye 2025).

[242]This figure was derived from architectural analysis rather than official disclosures (see Emberson 2025).

Much of it was focused on AI infrastructure such as chips, servers, and data centers.[243]

The specific numbers tell the story. Amazon reportedly spent over $100 billion in CapEx overall, dedicating most of it to AI infrastructure, including dozens of data centers globally for AI workloads, cloud computing, and edge services that sought to bring computing power closer to end users to reduce latency (Amazon Inc. 2025). For example, the company invested heavily in Project Rainier, a cluster of AI data centers outfitted with hundreds of thousands of its powerful Trainium2 chips, which it custom-designed to train AI models (Albergotti 2024; Amazon Inc. 2024). Meta's CapEx was $64–$72 billion (Meta Platforms Inc. 2025a). For instance, the company announced a new data center campus in Bowling Green, Ohio totaling 715,000 square feet and representing an investment of more than $800 million (Meta Platforms Inc. 2025b). Alphabet spent $75 billion on CapEx, including billions to both build out its data centers and train its generative AI models (Alphabet Inc. 2025). Meanwhile, Microsoft invested over $80 billion on AI data centers (Smith 2025). For example, it started building a $3.3 billion

[243] The scale of this investment wave strained traditional corporate financing mechanisms. In September and October of 2025 alone, Big Tech companies raised $75 billion through investment-grade bonds specifically for AI data center development, with banks syndicating an additional $38 billion in loans tied to Oracle and Vantage facilities—dwarfing these firms' pre-pandemic annual average of $37 billion in total bond issuance (Raitano and Cooper 2025). The financing mechanisms grew increasingly creative; Meta, for instance, structured its $60 billion AI infrastructure raise so that half would not appear as debt on its balance sheet, employing sophisticated financial engineering to do so.

data center in Mount Pleasant, Wisconsin—which was expected to be completed in 2026 (Mims 2025b).[244]

Established data center firms were drawn into the AI buildout as well. To give just one example: Iron Mountain, a long-time leader in enterprise and archival data management, broke ground in 2025 on its first Miami facility—marking the company's entrance into the AI infrastructure race. The company's data center was slated to deliver 16 megawatts of power capacity and 150,000 square feet of state-of-the-art cloud infrastructure to meet surging cloud AI demand (DataCenterDynamics 2025).

CapEx Numbers for Startups Were Also Mind Blowing

The 2025 investment frenzy extended well beyond public giants to the startup ecosystem, where OpenAI effectively redefined the meaning of scale. To put their ambitions in perspective, consider that GPT-4's final training run reportedly required roughly 25,000 NVIDIA A100 GPUs running for roughly three months (Parker et al. 2025). Against that baseline, the "Stargate" initiative—a collaboration between OpenAI, Oracle, and SoftBank—represented a staggering leap in magnitude. Boasting a projected price tag of $500 billion, Stargate targeted an initial 5 gigawatts (GW) of capacity with aspirations to reach a 10 GW build-out over four years (OpenAI 2025h). By December 2025, the project was accelerating ahead of its original 2026 schedule, having confirmed five specific sites across Texas, New Mexico, Ohio, and Wisconsin with a revised initial target of 7 gigawatts. Simultaneously, the company expanded its footprint globally: OpenAI and its partners announced "Stargate UAE," a 1-gigawatt compute cluster in

[244] I should also note that TSMC spent approximately $10 billion on CapEx in the second quarter of 2025 alone to support AI chip production (Mims 2025b).

Abu Dhabi intended to anchor a 5-gigawatt UAE–U.S. AI Campus, with the first 200-megawatt tranche expected to go live in 2026 (Nellis 2025).[245]

For its part, OpenAI's rival, xAI, raised over $10 billion in late 2024 and early 2025 to finance its own compute infrastructure, the aptly named "Colossus" supercomputer—a massive data center in Memphis designed to house hundreds of thousands of GPUs (NVIDIA 2024c; Wiggers 2024b; Jin et al. 2025). Following the activation of the original site, xAI broke ground in late 2025 on "Colossus 2" in the Greater Memphis area—a facility designed to push the company's aggregate compute capacity closer to the elusive one-million-GPU mark (Reuters 2025a).[246] To sustain this industrial scale buildout, xAI announced one of the largest private capital raises in tech history, numbering around $20 billion (Arroyo and Ludlow 2025; Jin et al. 2025).

Still other foundation model startups participated in 2025's AI-palooza. Anthropic made headlines in late 2025 with a major U.S. infrastructure commitment, announcing a $50 billion plan with Fluidstack to build custom data centers in Texas and New York, with sites coming online through 2026 (Anthropic 2025d). Perplexity also continued rapid expansion; by September 2025, the company had raised a cumulative total of approximately $1.5 billion in

[245] A separate partnership involving Abu Dhabi's ADQ and Energy Capital Partners began developing 25 GW of power capacity for data centers with a $25 billion investment (ADQ 2025).

[246] The original Colossus facility housed approximately 230,000 Nvidia GPUs, while Colossus 2 was slated to add another 550,000 units. Additionally, at the tail end of 2025 xAI purchased an 810,000-square-foot warehouse in Southaven, Mississippi, adjacent to Colossus 2, for conversion into a third data center powered by an onsite natural gas plant (Wayt 2025).

funding across several rounds, including a $200 million raise at a $20 billion valuation that year, using much of this capital to expand product capacity, partnerships, and the compute footprint required to serve growing query volumes (Temkin 2025; Field 2025b).

Finally, this era also witnessed massive funding rounds for other startups in more obscure segments of the AI stack (Reuters 2025b; Teare 2025). For example, Databricks received $10 billion during a Series J funding round (Databricks 2024); CoreWeave raised $1.1 billion in equity financing and secured $7.5 billion in debt financing (Blackstone 2024).

Consequently, venture capital firms raised unprecedented war chests during 2025 to keep the party going. For example, Lightspeed Venture Partners secured over $9 billion in the latter part of the year to fund the skyrocketing compute costs of its portfolio companies (Griffith 2025).

Looking Beyond Obvious AI Infrastructure Investments

During the early 2020s, other AI-related costs piled up, including investments in high-bandwidth, low-latency networking equipment and in cooling and environmental controls—supporting infrastructure that is essential for operating modern AI-optimized data centers at scale (IEA 2025). Furthermore, ongoing operational costs like electricity and maintenance, as well as efforts to offset hardware depreciation, added to the overall expense.[247]

Over this period, training AI models required massive parallel data throughput—the ability to move huge amounts of data at the same time across many GPUs. Consequently, the

[247] They also included the adoption of technologies like High Bandwidth Memory and advanced DDR5 memory, which offer significantly higher data transfer rates compared to their predecessors (Micron 2024).

demand for high-performance networking infrastructure—the hardware, software, and physical connections needed to link AI systems—surged. This was amplified by the growth of AI traffic, especially with bandwidth-intensive content that requires a lot of data, like video and high-resolution images (Washburn 2024). To meet this demand, companies made substantial investments in specialized network infrastructure to enable high-speed data transfer. This included spending on high-end switches and routers—the traffic controllers that direct data packets through networks—with increased bandwidth and reduced latency.[248]

This global demand for advanced networking also drove major investments in connectivity infrastructure during the early 2020s. Major tech companies expanded or announced new transoceanic submarine cable projects to boost international bandwidth and keep pace with growing data demand, including AI-driven services (Meredith 2025). Simultaneously, telecom carriers and ISPs continued to upgrade their backbone capacities (the main, high-capacity data routes of the internet) to handle exponentially growing AI traffic (Washburn 2023), implementing extensive fiber optic deployments and advancing toward higher data rates.

Substantial investments in AI infrastructure were accompanied by a sharp rise in energy use. While the electricity needed to power groups of thousands of AI interconnected chips rivaled traditional manufacturing plants, some of the larger training runs consumed as much energy as a small city (Strubell et al. 2019). Estimates for training large frontier language models suggest

[248] The most prominent example is Remote Direct Memory Access (RDMA)—a technology that allows computers in a network to access each other's memory without involving their operating systems—which allows data to travel over ethernet networks with very low latency and high throughput (Arista Networks 2024).

substantial energy use and emissions, though precise figures are difficult to verify for proprietary systems (see Patterson et al. 2021). By 2024, data center electricity consumption was estimated at approximately 415–416 terawatt-hours (TWh), accounting for around 1.5% of global electricity demand, with projections indicating it could reach 3% by 2030. (IEA 2025).[249]

Power hungry AI developers and hyperscalers therefore turned to nuclear power. As of mid-2025, several of them had forged agreements to secure this energy source for their data centers. This included Amazon's persistent pursuit of nuclear baseload, restructuring its agreement with Talen Energy to supply nearly 2 gigawatts from the Susquehanna plant after initially navigating complex regulatory hurdles (American Nuclear Society 2025).[250] Similarly, Meta signed a 20-year power purchase agreement for roughly 1.1 gigawatts of emissions-free output from Constellation's Clinton Clean Energy Center (Patel 2025). And Microsoft signed a 20-year PPA in 2024 with Constellation to provide power for data centers in the mid-Atlantic region from the mothballed Three Mile Island nuclear plant (Mandler 2024).

Beyond these deals around nuclear energy, many companies chose to bypass the conventional electric grid entirely by erecting their own on-site generation facilities to secure power faster and more directly amid supply-chain bottlenecks and permitting delays. OpenAI and Oracle installed natural gas-fired turbines for their $500 billion Stargate project in West

[249] The electricity component of U.S. CPI had already been growing faster than the overall index since 2022, coinciding with ChatGPT's launch (BLS 2025a).

[250] This expansion was made possible only after a clarifying order from FERC in December 2025 finally resolved the grid interconnection dispute that had stalled the project the previous year.

Texas, while xAI deployed gas turbines at its Colossus data centers in Memphis. Equinix implemented fuel cells across more than a dozen data centers nationwide (Hiller 2025).

During this same period, AI companies also made huge investments in advanced cooling solutions. This included innovative techniques such as direct liquid cooling (DLC), immersion cooling, and rear-door heat exchangers (RDHx). These investments appreciably improved the reliability and performance of AI hardware during both model training and inference (Vertiv 2025; Utility Dive 2024).[251]

Numerous smaller firms besides these usual suspects used their growing fortunes to gain traction on several problems. MLOps and ModelOps platforms like Databricks, Weights & Biases, and Hugging Face raised considerable capital to streamline model training, deployment,

[251] DLC involves circulating a liquid, such as water or a specialized coolant, directly over or through the hot components of the computer, much like how a car engine is cooled. This method is highly efficient because liquid transfers heat far more effectively than air. Taking liquid cooling a step further, immersion cooling fully submerges computer servers or individual components in a non-conductive liquid. This liquid draws heat away from the chips very efficiently, allowing for even denser and more powerful setups. Finally, RDHx are specialized doors placed at the back of server racks; instead of simply blowing hot air into the room, these doors contain coils with chilled water that absorb the heat directly from the servers as it exits. This prevents hot air from mixing with the cold air in the data center, making the overall cooling system significantly more efficient.

and monitoring (Bautista and Harrison 2023).[252] Databricks spent a small fortune to enhance its data management and machine learning platforms (Soper 2024). CoreWeave used its massive venture funding to expand its data center capacity and optimize its AI workload infrastructure (Coreweave 2024). Meanwhile, firms such as EdgeQ, Swim.ai, FogHorn Systems, ZEDEDA, and ADLINK Technology, built a bevy of smaller data-center nodes near end users to facilitate edge computing and IoT local AI inference (Savino 2024; Torchia and Shirer 2024).

These developments reveal AI's commercialization as more than technological change—it has reorganized firm boundaries, shifted bargaining relationships across the supply chain, and redirected capital allocation economywide. The infrastructure buildout documented in this chapter represents the largest private-sector capital wave since the late 1990s, with consequences that extend far beyond the tech sector—a topic I will explore further in the next chapter.

Circular Deals

Finally, a word of caution about the larger economic impact of the AI boom: much of the capital surging into AI infrastructure in 2025 was routed through increasingly complex, circular arrangements—a pattern with significant implications for how we interpret the investment figures disclosed above. Major firms—including OpenAI, Nvidia, Oracle, CoreWeave, and cloud hyperscalers—were engaging in cross-investments, vendor financing, and equity swaps that ultimately made it difficult to distinguish genuine capital formation from accounting artifacts.

---

[252] As the previous chapter outlined, MLOps focus on the reliable and efficient deployment, maintenance, and monitoring of machine learning models in production. ModelOps broaden this scope to include all model types, with a stronger emphasis on governance, comprehensive lifecycle management, and cross-functional collaboration.

For example, when Nvidia invests in CoreWeave, which uses the funds to buy Nvidia chips, which generates revenue that Nvidia reports as sales growth, the “investment” partially cancels itself out in economic terms even as it inflates reported figures.

This circularity echoed the dot-com era's vendor financing excesses, though at exponentially larger scale. During the late 1990s, telecom-equipment makers like Lucent lent billions to startup customers to purchase equipment, fueling apparent sales growth that evaporated when those customers failed. The AI era's version of circularity proved far more complex—involving equity investments, compute credits, and multi-party swaps rather than simple loans—but the fundamental risk remained similar (Weil 2025). Accordingly, following reports detailing the extent of these interdependencies—dubbed the "Trillion-Dollar Tangled Web"—investors began applying a "circularity discount" to pure-play infrastructure firms. Notably, secondary market valuations for firms like CoreWeave softened in the fourth quarter of 2025 as the market repriced revenue streams perceived as overly dependent on vendor-financed affiliates (see Forgash and Ghosh 2025).

Notwithstanding some creative accounting, these mega-projects signaled a definitive shift from an era of software experimentation to one of heavy industrial capitalization. This distinction is crucial for understanding the economic sequence of the AI revolution. While the next chapter will investigate the long-run effects of this technology on total factor productivity, labor markets, and employment, the economic story of the mid-2020s was defined by this short-to-medium-term construction boom. Before AI could "eat the world" as software did, it first had to consume

a staggering amount of energy, silicon, and steel.[253] Thus, the immediate economic impact of AI was not measured in jobs lost or gained, but in the sheer magnitude of fixed capital formation required to build the engine itself—a "demand shock" for physical infrastructure that constituted the primary economic event of the decade's midpoint.

Beyond Inputs in, Outputs out: Economizing on Scarce Factors

During the early 2020s, high inference costs strained even the largest tech balance sheets, while capital barriers locked smaller firms out of the AI race entirely. Researchers responded by experimenting with model architectures fine-tuned on smaller datasets, drastically reducing training time and the computational resources required for deployment.

Starting with DeepSeek, several providers matched OpenAI's quality at a fraction of the cost, commoditizing what had recently been premium features (Urbain and Lin 2025). In turn, incumbents adjusted their model development strategies. For example, OpenAI began tapping

---

[253] Marc Andreessen, the famed venture capitalist who founded Netscape, argued that software companies were poised to take over large swathes of the economy by digitizing value chains in industries ranging from movies to agriculture (Andreessen 2011). The era he described was characterized by low marginal costs and asset-light business models—startups could disrupt incumbents with little more than code and cloud credits. The AI boom of the mid-2020s represents a structural inversion of this thesis (see Schrage and Kiron 2025). While AI may eventually "eat" the economy in terms of automation, at the time of this writing its production phase remained remarkably "industrial"; as documented above, massive fixed capital investment in energy, semiconductors, and real estate mirror the steel and rail booms of the 19th century rather than the software boom of the 2010s.

TPUs rented through Google Cloud to conduct inference at a lower per-unit price than NVIDIA's high-end GPUs (Gardizy and Liu 2025). Simultaneously, as user volumes surged, providers sought economies of scale through batching—processing multiple requests simultaneously to maximize hardware utilization (The Information Staff, 2024). These hardware efficiencies proceeded alongside massive software improvements; MLOps platforms and automated data annotation tools matured just in time to bring inference costs down.

This phenomenon is consistent with earlier industrial revolutions: innovations that economize on scarce factors unlock adoption even when capital outlays remain high. Classic General Purpose Technology diffusion theory anticipates a three-phase trajectory—initial novelty, a consolidation lag, then takeoff once complements lower effective use costs and broaden applicability (Helpman and Trajtenberg 1996; Bresnahan 2010). In AI, that pattern manifests in three coupled ways: (i) rapid declines in the marginal cost of inference, (ii) maturation of complements (tools, APIs, orchestration, governance) that let organizations embed models into workflows, and (iii) equally rapid increases in usage. I consider each in turn.

OpenAI's posted API list prices for GPT-4-class models fell sharply as model offerings evolved and list prices were revised.[254] At GPT-4's launch in March 2023, Nebuly (2024) reports list prices of $30 per million input tokens and $60 per million output tokens (i.e., $0.03 and $0.06 per 1,000 tokens, respectively). On the output side alone, these figures imply a drop from $0.06 per 1,000 output tokens (GPT-4 launch) to $0.01 per 1,000 output tokens (later GPT-4o

[254] While not a direct observation of providers' underlying marginal production cost, these list prices are best interpreted as the published per-token charge for running the model and exclude fixed costs such as training and R&D (see OpenAI 2025d; Gundlach et al. 2025).

pricing), an output-price reduction of roughly 83%. OpenAI announced GPT-4o in May 2024 at $5 per million input tokens and $15 per million output tokens ($0.005 and $0.015 per 1,000), and subsequent GPT-4o snapshots were priced lower (e.g., $2.50 per million input and $10 per million output). Taken together, these posted prices document a rapid compression in the metered unit price of inference for flagship models (separately for inputs and outputs).

Meanwhile, Figure 10.3 graphs the number of web visitors to the ChatGPT chatbot. The plotted line (three-month moving average; logarithmic scale) captures a dramatic expansion of traffic, growing nearly two orders of magnitude over only two years.

As expected, this figure captures the striking three-phase trajectory typical for General Purpose Technologies (see Agrawal et al. 2023). During Phase 1 ("novelty era"), from its November 2022 debut through spring 2023, visits surged to roughly 170 million monthly sessions. During Phase 2 in mid-2023 ("consolidation"), the initial hype cooled and traffic plateaued near 130–140 million. During Phase 3, from late-2023 onward ("reacceleration"): fueled by product maturation and the workflow integrations described earlier, adoption skyrocketed once again. Monthly visits climbed past 1.8 billion by March 2024, doubled to 3.9 billion by early 2025, and reached 6.3 billion by September 2025—cementing ChatGPT's transition from a niche tool to a utility-scale platform.

Figure 10.3. Exponential Adoption Pattern of Web Visitors to ChatGPT Chatbot.

[INSERT FIGURE 10.3]

Notes: This figure depicts global monthly web visits to ChatGPT's public interface between November 2022 and November 2025. A "visit" represents a session, not a unique user; totals should not be interpreted as monthly active users. Mobile app usage is excluded. The series spans November 2022 through November 2025, aggregating traffic across both chat.openai.com

and chatgpt.com following OpenAI's domain migration in May 2024. To reduce single-month volatility and highlight medium-term trends, the plotted line applies a three-month moving average on data transformed into natural logarithms; tick marks labeled in billions. Source: monthly Similarweb traffic reports, including Carr (2024); Barenholtz (2025); Scott (2025) and Similarweb (2025).

This reacceleration phase coincided with the steep decline in marginal inference costs. As token prices collapsed by over 80%, the feasible design space for AI applications expanded dramatically, enabling a qualitative shift in how models were utilized. At premium per-token prices, AI was largely confined to single-prompt questions and answers. However, as costs approached near-commodity levels, users could use chatbots for more goal-directed, multi-step processes that involved planning, retrieving, summoning tools, and iterating, helping to explain the increased adoption rate mapped by Figure 10.3. While these complex workflows consume far more tokens per completed task, thanks to the epic collapse in AI inference prices I outlined above, they became economically feasible at scale. In short, falling costs didn't just save money; they fundamentally altered how this technology could be used.

In turn, this widespread validation of the technology's utility provided the necessary signal for the ecosystem to mature. With the demand side proven, firms had clear incentives to industrialize deployment: to continue to lower unit costs, increase reliability, and embed models into existing workflows. Therefore, standardization and orchestration efforts ensued across model formats, benchmarking suites, deployment tooling, and governance interfaces.

EARLY AI STANDARDIZATION

AI standardization proved thoroughgoing, quickly transforming four distinct, reinforcing layers: foundational interoperability, operational processes, application access, and societal governance.

Before exploring how progress towards standardization proceeded across each of these layers in depth, allow me to briefly review the political economy of technological standardization.

The economic logic underpinning AI's first standardization wave can be understood through the lens of transaction costs—the frictions that impede exchange and coordination across organizational boundaries. Before standardization, every layer of the AI stack imposed its own barriers: firms seeking to adopt AI faced steep access costs (acquiring specialized talent and infrastructure), high transaction costs (negotiating bespoke integrations, managing vendor lock-in, navigating incompatible systems), and substantial development costs (building custom pipelines from scratch rather than assembling proven components). These frictions did not merely slow adoption; they concentrated AI capabilities in the hands of well-resourced incumbents who could afford to absorb them.

Standardization attacks each of these barriers simultaneously. Common formats and protocols reduce transaction costs by enabling arms-length exchange: a model trained by one organization can be deployed by another without costly renegotiation or custom engineering. Shared infrastructure and tooling reduce development costs by converting one-time investments into reusable public goods: a pipeline built to one standard can be replicated across firms and industries. And simplified interfaces such as no-code platforms reduce access costs by abstracting away complexity, allowing firms without dedicated AI teams to participate in the ecosystem.

The cumulative effect is to transform AI from a bespoke capability available only to frontier firms into a commodity input accessible throughout the economy. This is the economic mechanism through which a general-purpose technology achieves broad diffusion: not through

the technology's raw capability alone, but through the institutional and technical infrastructure that reduces the cost of deploying that capability at scale.

However, as with any process that improves efficiency and promises a greater producer surplus in the process, conflicts over the distribution of the spoils may arrive. Previous industrial revolutions have demonstrated that resolving distributional conflict may appreciably delay or derail joint efforts to shift demand curves (Haber and Lamoreaux 2021). Therefore, as Chapter 3 outlined, institutions and norms often arise to reduce frictions and solve coordination problems that abide by the logic of the "Battle of the Sexes"—relative winners and losers who jostle to acquire a greater share of the rents.

In what follows, I trace the early AI standardization process across four reinforcing layers—foundational interoperability, operational processes, application access, and societal governance—showing how each contributed to the democratization of AI. I then transition into speculating about what future standardization efforts may look like and their potential effect on the diffusion of AI.

The Foundational Layer: Interoperability of Core Components

Before AI could be widely integrated, its essential building blocks needed to work together seamlessly, regardless of who made them. At the most basic level, uniform media and data formats underpinned interoperability. Standardized image types like JPEG and PNG, along with common data schemas such as CSV, underwrote applications from medical-image analysis to autonomous-vehicle vision (Daffodil Software 2025; Madden 2025). Battle-tested open-source libraries further smoothed model training and deployment, fostering collaboration and knowledge sharing (Alok 2024).

Building on this, the Open Neural Network Exchange (ONNX) served as the "Rosetta Stone" for the AI industry. Previously, models were essentially trapped in the frameworks where they were created; moving a model from PyTorch to TensorFlow, or getting it to run on a specific microchip, was a technical nightmare requiring complex, error-prone translation. ONNX solved this friction by establishing a universal standard for how models are saved and shared, allowing a system built in one language to run natively on any compatible machine (Chia 2024).

By decoupling software from hardware, ONNX allowed chipmakers to compete purely on performance rather than ecosystem lock-in. Crucially, this standard also triggered a vital second-order effect: it made possible the rise of massive, vendor-neutral model libraries like Hugging Face (Bonnet 2024). Before such repositories existed, accessing a pre-trained model meant tracking down the original research team, hoping they had released their weights, and wrestling with idiosyncratic code dependencies—a process that created enormous friction for developers and researchers alike. Once models could be saved in a universal format, however, centralized hubs became viable. Model libraries provided a single destination where thousands of pre-trained models could be discovered, downloaded, and deployed through a standardized interface, regardless of who trained them or on what hardware.

This dynamic transformed AI development: instead of building from scratch, developers could now browse a catalog of ready-made capabilities—text classification, image segmentation, translation, summarization—and integrate them into applications with a few lines of code.

With portability solved, the market required a transparent way to measure value. This gap was filled by MLPerf, a standardized suite of benchmarks that functioned effectively as a consumer protection standard for the AI era (Mattson et al. 2020). Instead of relying on vendor-supplied marketing claims, MLPerf established rigid, independent tests for core tasks—

ranging from image classification to language translation—using fixed datasets and accuracy targets (MLCommons 2024).

By enforcing an apples-to-apples comparison, the standard forced hardware vendors to compete on objective metrics like performance-per-watt. This created a "race to the top" that ultimately slashed the energy and capital costs of large-scale AI workloads, further accelerating the technology's commoditization (Huescas 2024).

The Operational Layer: Standardizing the "AI Factory Floor"

With interoperable components in hand, the next challenge was standardizing the process of assembling them into reliable, production-grade systems. Before this maturation, deploying an AI model at scale was often artisanal work: each company cobbled together bespoke pipelines for data preparation, training, and serving, with limited transferability between projects or organizations.

Kubernetes emerged as a general-purpose "operating system" for containerized workloads, managing underlying compute infrastructure in a largely vendor-neutral way (Kubernetes 2024). Building on this foundation, Kubeflow standardized AI-specific workflows—providing reusable modules for data preprocessing, distributed training, and model serving that can be deployed across Kubernetes environments (Kubeflow 2024). This shifted AI deployment from a craft toward a more industrial process, reducing the switching costs of operationalizing similar ML workflows across teams and environments (Kubeflow 2024).

In parallel, the discipline of MLOps matured from a niche skill into a standardized operational process. Previously, managing the lifecycle of production models was often ad hoc: experiments were tracked inconsistently, versioning was uneven, and deploying updates required manual steps across the stack. Tools like MLflow and Kubeflow standardized core functions—

experiment tracking, version control, automated deployment, and monitoring—into components organizations could adopt without rebuilding the workflow from scratch (MLflow 2024; Kubeflow 2024). By 2025, an MLOps pipeline could, for instance, be configured to retrain a fraud-detection model on a regular cadence using fresh transaction data, validate performance against defined benchmarks, and deploy updates with limited manual intervention (Shakudo 2025).

Equally important was the commercialization of techniques that made AI behavior more predictable. Reinforcement learning from human feedback (RLHF) emerged as a critical technique for aligning model behavior with user expectations, but early implementations often required bespoke pipelines, including custom reward models and feedback loops. By 2025, however, important elements of RLHF were increasingly "productized" via cloud and open-source pipelines that provided reusable components for preference-data handling, reward modeling, and reinforcement stages (Mahaseth 2025).[255] This standardization helped make AI outputs more consistent—supporting interoperable, plug-and-play integration across enterprise systems. In telecommunications, for example, Verizon deployed AI-driven personalization through its myPlan feature, using machine learning to support tailored recommendations and offers (Wassel 2024).

The Access and Application Layer: Simplifying Integration and Use

Once AI could be built and deployed reliably, the focus shifted to distribution: making this intelligence easy for developers and users to consume. Previously, integrating AI into an

[255] Even so, many real-world RLHF implementations remained bespoke, with substantial custom work in data collection, evaluation design, and governance.

application meant wrestling with model weights, configuring inference servers, and provisioning GPU clusters—a process that required dedicated machine learning engineers and significant capital investment (Merizzi et al. 2025).

The API economy for AI changed this calculus entirely. APIs function as the "messengers" of the software stack. They enable a client application to send a simple request (e.g., "classify this text") and receive a structured result, while completely abstracting away the immense complexity of the underlying "AI factory." By treating intelligence as a utility call—similar to fetching weather data or processing a credit card—this layer democratized access, allowing non-specialist firms to integrate state-of-the-art AI without hiring a single machine learning engineer.

Specifically, platforms like Google Cloud AI, Amazon Bedrock, and Azure OpenAI wrapped complex models in simple, high-level endpoints, allowing developers to implement capabilities like chatbots, sentiment analysis, or translation with a single line of code (Google Cloud 2025b; Disu 2024; Nanda 2025). Hugging Face complemented its model repository function with hosted inference endpoints, further lowering the barrier for developers who wanted to experiment before committing to enterprise platforms (Hugging Face 2025a).

Simultaneously, standardization extended beyond the cloud to the "edge"—the devices in our pockets. Previously, developers seeking to run AI locally faced a fragmented landscape: they could attempt to port open-source models to mobile hardware, but this required extensive optimization work and deep expertise in on-device inference. In 2025, Apple introduced the

Foundation Models framework, a developer toolkit that embedded Apple's own Large Language Models directly into the iOS operating system.[256]

Unlike frontier-scale models running in data centers, Apple's on-device foundation model is deliberately smaller and optimized for fast, privacy-preserving text tasks and structured output directly on user hardware. This smaller size allowed them to tackle "atomic" micro-tasks—inline text suggestions, tone analysis, and placeholder generation—extremely quickly on the device's silicon. The primary audience was mobile developers who wanted to layer AI capabilities into their apps without incurring the latency or per-call fees of cloud APIs, while simultaneously preserving user privacy.

By providing a stable interface and a ready-made Software Development Kit, Apple obviated the need for developers to write complex inference code. Rather than routing every prompt through a remote server, apps could invoke these on-device models at zero marginal cost. Google offered a comparable capability with its small Gemma models for Android, but Apple's approach was more deeply integrated: where Android developers had to bundle model files inside their apps—inflating file sizes and requiring updates whenever the model improved—Apple's models came preloaded in the OS, with improvements arriving automatically via system updates.[257]

---

[256] This passage draws strongly on Holmes and Tilley (2025).

[257] Google's Gemma models ranged from 1–3 billion parameters and were offered free to Android developers. The technical limitation was the Android Package Kit (APK) architecture, which required bundling model files within the app itself rather than accessing them at the OS level.

This development marked the final link in the chain: by 2025, standardization spanned from the gigawatt-scale data centers of the "AI Factory" all the way to the smartphone in a user's pocket—extending plug-and-play AI to every layer of the stack.

In parallel, the challenge of connecting these models to the outside world was addressed by Anthropic's Model Context Protocol (MCP), which rapidly established itself as the "universal translator" for AI connectivity (Anthropic 2024). Prior to MCP, connecting an LLM to a corporate database or a tool like Slack required writing bespoke "glue code" for every single integration—a high-friction process that siloed intelligence (Palazzolo 2025). MCP solved this by defining a standardized protocol for how AI models discover and interact with external resources. Instead of building custom integrations for each tool, developers could write once to the MCP standard and connect to any compliant system instantly (Anthropic 2024).

Building upon this interoperability, a thriving secondary market for plugins and integrations emerged by 2025. Previously, extending an AI system's capabilities required professional developers who could navigate APIs, manage authentication, and write integration code. The new generation of standardized protocols and "zero-code" interfaces lowered this barrier dramatically. A new layer of startups—tool library platforms like Speakeasy and Stainless—emerged to help companies build MCP-compatible interfaces, enabling agents to interact with their services (Sullivan 2025; Stainless 2025; CB Insights 2025c).

In turn, adoption barriers plummeted. Citizen developers—everyday users without formal programming training—could now assemble reusable components like payment processors or analytics dashboards using simple text commands and community-shared templates. What had once required a software engineering team became a drag-and-drop exercise.

These advances—cloud APIs, edge deployment, and protocol standardization—soon enabled a new class of application: the AI agent. Unlike static models that responded to single prompts, agents could autonomously execute multi-step workflows, navigate software interfaces, and make decisions over extended time horizons.

But this capability introduced a fresh coordination problem. If agents from different vendors needed to collaborate—or if a single agent needed to interact with dozens of enterprise tools—the absence of shared conventions for discovery, communication, and handoff threatened to recreate the fragmentation that earlier standardization efforts had overcome. While I return to this challenge, and the industry's nascent response to it, further below, I now examine the final layer of the standardization stack: governance and trust.

The Governance and Trust Layer: Enabling Broad Societal Adoption

Beyond technical interoperability and agent coordination, widespread adoption faced a final, formidable hurdle: trust. For AI to permeate high-stakes sectors like finance and healthcare, it had to earn institutional confidence regarding data privacy and safety.

Privacy-preserving collaboration—long considered the "holy grail" for regulated industries—was finally operationalized through the standardization of Federated Learning protocols.

Frameworks like FATE and TensorFlow Federated (TFF) established a standard for secure, encrypted exchanges of model updates. This allowed organizations to "co-train" a shared AI model on siloed data without ever sharing the raw customer records themselves (Stuecke 2025). By standardizing the protocol for *how* models learned from private data, competitors could collaborate on fraud detection or medical research without exposing their proprietary assets.

As AI matured into regulated sectors, formal governance standards took hold. Driven by both governmental initiatives and the collaborative efforts of Standard Setting Organizations (SSOs), frameworks emerged to proactively address key risks. NIST's AI Risk Management Framework laid out methods to identify and mitigate bias, security flaws, and opaque decision paths (Miller 2024), while an international committee ratified global guidelines on AI explainability, robustness, accountability, and ethics (Diab and Mullane 2024).[258]

As Chapter 12 will document, a prime example of this proactive governmental push was the European Union's AI Act. In practice, the Act replaces vague legal debates with a clear "presumption of conformity": if a company follows harmonized technical standards—essentially a government-approved checklist for safety and accuracy—regulators assume the system is legal. This transforms compliance from a bespoke, expensive legal battle into a standardized engineering task, allowing vendors to "certify once" and deploy systems across different sectors.

Accelerated Adoption of AI Applications

The cumulative effect of this four-layer standardization—foundational interoperability, operational processes, application access, and governance frameworks—was to systematically

[258] More precisely, the international subcommittee on AI that was jointly established by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) under their Joint Technical Committee 1 (JTC 1) on Information Technology. It is responsible for developing global standards for AI, including around trustworthiness, governance, and ethical use. The ISO develops and publishes international standards across industries. The IEC focuses on standards for electrical and electronic technologies (see ISO/IEC JTC 1/SC 42 2024).

dismantle the barriers that had confined AI to well-resourced incumbents. Access costs fell as APIs and SDKs abstracted away the need for specialized talent. Transaction costs fell as common protocols like ONNX and MCP enabled arms-length exchange without bespoke negotiation. Development costs fell as reusable pipelines and pre-trained models converted one-time investments into shared infrastructure. And compliance costs fell as certification frameworks allowed vendors to demonstrate trustworthiness through standardized audits rather than case-by-case legal review.

The practical ramifications were dramatic. Tasks that once demanded specialized teams, lengthy integration cycles, and heavy compliance overhead were now achievable by small, cross-functional teams in a fraction of the time—thanks to the compounding effect of collapsing costs across the code, compute, and organizational layers.

HOW TO MAKE SENSE OF AI AS A STILL MATURING GPT

Bick et al. (2025) show that, as of late 2024, nearly 40 percent of U.S. adults aged 18-64 used generative AI, with 23 percent of employed respondents having used it for work at least once in the previous week and 9 percent using it every workday. And, when measured from each technology's first mass-market product launch, generative AI's workplace adoption in 2024 exceeded the pace of personal computers in 1984, with 32% of workers using generative AI compared to 25% for PCs, while its overall adoption (45%) surpassed both PCs (20%) and the commercial internet (21%).

Yet, despite this seemingly widespread adoption pattern, most organizations that experimented with ever cheaper AI during the early 2020s limited themselves to tentative experiments in lower risk domains that stopped short of fully automating workflows. Challapally et al. (2025) document that, even as CEOs remained publicly bullish about AI’s potential, 95% of

corporate AI pilots were not delivering meaningful value. And whatever gains did materialize, they were often unglamorous: summarizing documents, automating customer service calls, and extracting data from PDFs and emails. As one group of technology journalists observed, AI in business was "ahead of where people generally think it is, but behind where they imagine it could or should be" (Bousquette et al. 2025).

This gap between capability and deployment raised a puzzle: why wasn't the emerging "diffusion infrastructure" outlined above doing more to reduce the frictions of AI adoption across enterprises?

Just as General-Purpose Technology theory helps explain the viral adoption figures expressed in Figure 10.3, it also helps make sense of the apparently lackluster diffusion of AI circa 2025. The answer lies in the distinction between adoption and transformation. As Brynjolfsson et al. (2021) note, organizations initially use new technology merely to substitute for the old—using AI to write code or emails faster—rather than reimagining the workflow itself. Indeed, as discussed in Chapter 2 of this book, David (1990) famously documented that during the electrification of factories, a first wave merely replaced steam engines with electric motors without altering the fundamental architecture of production.

For wider and deeper technological diffusion to take place, two things must happen: first, a deeper, more thoroughgoing wave of standardization; and second, the accumulation of "intangible capital" in the form of new business processes, reskilled workforces, and novel organizational structures (Bresnahan and Greenstein 1996). It took deeper technological standardization for electricity to be reliable, cheap, effective, and safe enough for factories to reinvent themselves to better exploit the new power source. Specifically, the widespread adoption of electricity necessitated agreements on voltage, currents, and socket designs (see

Chapter 2). Similarly, the AI standardization efforts described earlier in this chapter solved the technical problem of *access*, but they did not yet solve the organizational problem of *redesign*.

Therefore, the failure of early AI pilots was not a failure of technology, but a failure of organizational architecture—and only a new, deeper round of institutional standardization alongside the right kind of organizational adaptation can truly close this gap. Until firms standardize both the interfaces between AI agents and enterprise data and the governance of autonomous decision-making the technology will remain trapped in the "pilot purgatory" of marginal efficiency gains rather than driving the promised revolution in value creation.

To gain purchase on how AI may yet fulfill its potential as a groundbreaking General-Purpose Technology on par with those behind other industrial revolutions, this subsection explores how AI's proliferation beyond 2025 may depend on even more intensive standardization and organizational adaptation efforts.

AI Standardization: Going Forward

Building on the established protocols outlined further above, the next wave of AI standardization may focus less on getting models to simply run, and more on ensuring they run safely, collaboratively, efficiently, and accountably. The path forward may involve creating better protocols for how AI systems interact, how their data is managed, how their safety is assured, and how their economic value is measured. Perhaps the most enticing possibility is the quest to make AI agents and other embodied AI applications fully interoperable.

The Fraught Quest for Agentic AI Interoperability

As outlined earlier, the AI agent space is bedeviled by a classic coordination problem. Agents from different vendors often needed to collaborate; similarly, a single agent usually needs to interact with dozens of enterprise tools. However, circa 2025, there was still a marked absence of

shared conventions for discovery, communication, and handoff to facilitate these exchanges. The industry faces a choice: either agents will remain Balkanized within proprietary ecosystems or a new layer of open infrastructure will need to emerge.

A common AI agent "language" would allow these agents to discover each other's capabilities, issue standardized task orders, bid on jobs, and verify completion—essentially creating a SWIFT-like network for AI agents. Consider that SWIFT (the Society for Worldwide Interbank Financial Telecommunication) is a global messaging network that financial institutions use to send and receive information about financial transactions in a secure, standardized, and reliable environment. It does not actually hold funds or execute transfers itself; rather, it provides the secure communication protocol that enables banks worldwide to communicate instructions for money transfers, foreign exchange, and other financial operations. The idea is that a similarly secure, standardized messaging layer for AI agents from different developers or organizations would allow them to interact without a central authority directly controlling the funds or performing the actions like how SWIFT enables global financial flows by standardizing interbank communication.

By late 2025, a major logjam blocking the dream of AI agent interoperability seemed to be removed. The major AI laboratories had arrived at a surprising conclusion: collaboration on open infrastructure could be more valuable than proprietary control. OpenAI, Anthropic, and Block co-founded the Agentic AI Foundation (AAIF), with AWS, Microsoft, Google, Bloomberg, and Cloudflare joining as platinum members (OpenAI 2025c; Block 2025). Each founding member contributed a key project to the foundation: Anthropic donated the MCP standard explained above; OpenAI contributed AGENTS.md, an open format for providing AI

coding agents with project-specific instructions (already adopted by over 60,000 open-source projects); and Block contributed goose, an open-source agent framework (InfoQ 2025).

The AAIF represented a strategic inflection point. Where earlier phases of AI development had been characterized by proprietary competition—closed model weights, exclusive cloud partnerships, vertically integrated stacks—the shift to agentic AI prompted a different calculus. AI agents that could autonomously navigate software and execute workflows required interoperability standards that no single firm could impose unilaterally. The similarity to earlier standards battles documented in this book is instructive: just as the internet's growth depended on open protocols like TCP/IP and HTTP, and just as smartphone ecosystem players pursued uniform cellular telecommunications protocols through SSOs, the proliferation of AI agents will depend on shared conventions for how digital workers discover tools, access context, and coordinate across platforms.

As in those earlier collective action problems, the response to this coordination challenge revealed that standardization is never purely technical; it is always also political. As Anthropic's product team noted, "the company defining how an industry works often captures more value than the company trying to own it outright" (Nuñez 2025).

By defining the standards for how agents talk to software, these model developers effectively positioned themselves as the architects of the new digital labor market. Their strategic goal was clear: ensure that the "intelligence" (the model) remained the scarce, high-value component, while the connections to data and tools (the protocol) became commoditized and frictionless. Indeed, control over interoperability and standards has long been recognized as a strategic lever for capturing value in networked markets (Shapiro and Varian 1999)— and

predictably, these standardization efforts soon sparked a fierce counter-reaction from incumbents attempting to protect their turf.

As AI agents grew more capable, enterprise software incumbents began limiting API access. Salesforce aggressively throttled its Slack API, effectively prohibiting non-partner apps from reading message history at scale (Slack Developers 2025). Similarly, Atlassian—the gatekeeper of the modern corporate workflow through its Jira (project tracking) and Confluence (knowledge base) platforms—overhauled its access rules to protect its territory. Recognizing that third-party AI agents needed to ingest thousands of project tickets and documents to function, Atlassian introduced a complex "points-based" limiting system that specifically targeted these high-frequency machine queries (Braun 2025).

These restrictions were not merely technical housekeeping; they were defensive maneuvers in a struggle for rent extraction. By locking down their APIs, incumbents sought to prevent their proprietary workflows from becoming mere "food" for third-party agents, insisting instead that the value creation—and the revenue—remain captured within their own interfaces.

Complicating this landscape further, even as they collaborated through the AAIF, important players in the AI space simultaneously pursued parallel standards of their own. Google released the Agent-to-Agent Protocol, an open standard enabling agents built on different frameworks to collaborate, drawing early contributions from major professional services firms like Accenture (Surapaneni et al. 2025). IBM released its Agent Communication Protocol for inter-agent messaging within enterprise environments (IBM Research 2025). And, in late 2025, Snowflake announced the Open Semantic Interchange initiative—an alliance with more than a dozen providers—to create a vendor-neutral standard for semantic metadata so AI agents can interpret and exchange business data across applications (Parbel 2025; Snowflake.help 2025).

Unlike the AAIF, which was led by model developers, this coalition represented data infrastructure players staking their own claim to define the terms of agent access.

Whether value would ultimately accrue to the standard-setters or the gatekeepers remained an open question—one whose resolution would shape the distribution of rents across the AI economy for years to come.

Promise of AI Agent Interoperability

What we can know with greater certainty, however, is that a more thoroughgoing standardization of data format and AI interfaces could unleash a mature AI agent economy that creates far greater value (Tan and Huang 2025). In this world, AI agents would be able to seamlessly pull live, relevant data from various internal systems and generate actionable content or insights directly within a user's workflow, proceeding to reach out to and interact with their peers in the digital world. By reducing the transaction costs of digital exchange, such protocols would transform isolated "walled gardens" of intelligence into a liquid, interoperable market defined by a division of labor defined by specialization along the lines of comparative advantage. Cutting out the human middle person would, not only be more efficient, but might generate unforeseen economic opportunities.

Imagine, for example, a world where AI travel agents could automatically contract with specialized local restaurant AI booking agents, or AI corporate logistics agents could hire independent AI trucking agents for specific deliveries. Building on the MCP standard established in 2025, these agents may leverage an MCP server—as I discussed above, these are essentially specialized software programs that act as intermediaries, making their own distinct functions or data readily available for AI models to use—to browse the web, query a database like Salesforce to retrieve client records, invite other agents to join a Slack channel, and respond to a query like:

"Summarize the recent activity for ‘Acme Corp’ and draft a follow-up email to their account manager with the latest product update" (Palazzolo 2025).

The same could hold true for embodied AI. While circa 2025, discrepancies in friction, mass distribution, and sensor noise made robots irreconcilable, standardized “world models” would allow skills learned in one simulator to transfer seamlessly across diverse robotic systems *sans* expensive, bespoke fine-tuning for every specific robot chassis (Wu et al. 2022). A standardized schema would function effectively as a "hardware abstraction layer" for the physical world, decoupling the AI’s decision-making from the specific mechanics of the robot—much like how standard device drivers allow an operating system to control any printer regardless of its internal engineering. In turn, standardization would drastically reduce development costs for applications in manufacturing, logistics, and home assistance (Sakagami et al. 2023). For instance, a robot designed to assist in a standardized hospital environment could instantly adapt its learned navigation and object manipulation skills to a newly designed wing, or even a different hospital entirely, without extensive reprogramming.

Getting Down to Brass Tacks

For all this to happen, standardization may need to further transform the core model and data layer, moving beyond basic interoperability to address increasing complexity and diverse data consumption (Kiela et al. 2021). Moving past the siloed formats of the early 2020s—where robotics data was trapped in proprietary formats like ROS bags and web data in text-heavy standards like JSONL—the emergence of an Open Multimodal Format (OMF) could serve as a

universal container for different data types.[259] By wrapping text, images, audio, and video with synchronized timestamps and metadata, this standard could revolutionize industries from media to autonomous systems—allowing car companies to share standardized log files containing video, LiDAR, and audio data (see Sakagami et al. 2023), or enabling media companies to process entire archives with any OMF-compliant model.

Additionally, technical standards for data lineage may become critical industrial infrastructure to prevent "model collapse"—the mathematical degradation that occurs when models train recursively on their own synthetic output (Shumailov et al. 2023). Just as C2PA (Coalition for Content Provenance and Authenticity) began establishing technical standards for media origins in the early 2020s (C2PA 2023), the next wave of standardization may encode invisible watermarks and metadata labels at the dataset level, distinguishing synthetic from organic tokens (Kirchenbauer et al. 2023). This goes beyond mere IP compliance; it creates a standardized "quality control" layer for the global data supply chain, ensuring that future foundation models can filter their training inputs to maintain performance and preventing the poisoning of the digital commons.

---

[259] ROS bags (Robot Operating System bags) are the industry-standard binary file format for recording time-synchronized sensor data from robots, such as LiDAR scans and camera feeds. JSONL (JSON Lines), by contrast, is a lightweight text-based format favored for training Large Language Models, where each line represents a distinct data record. In the early 2020s, these formats were fundamentally incompatible: a researcher could not train a language model on robotic experience without writing complex, bespoke scripts to extract and convert the heavy, hardware-centric sensor logs into a format the text-based model could understand.

With reliable data inputs secured, the focus shifts to hardening model resilience. A "SafetyPerf" benchmark suite, analogous to MLPerf—the industry standard for speed and throughput outlined further above—could provide the necessary "stress tests" for model reliability. Rather than vague promises of safety, this standard would quantify technical failure modes: measuring a model's specific vulnerability to prompt injection (security), its rate of hallucination per thousand tokens (reliability), and its drift under adversarial conditions (robustness).

However, these technical metrics are useless if they remain hidden in a lab. To unlock the next phase of enterprise adoption, the market may coalesce around a standardized disclosure framework akin to the Material Safety Data Sheets (MSDS) used in the chemical industry (OSHA 2012). Currently, procurement officers in regulated sectors are often paralyzed by information asymmetry: they cannot inspect the “black box” of a model to verify its training data or safety profiles. Building on the model-documentation framework proposed by Mitchell et al. (2019), a procurement-oriented Model Fact Sheet (MFS) could add machine-readable fields for training data provenance, intended use constraints, known failure modes, and environmental footprint, enabling automated enterprise risk assessment and vendor comparisons. Just as the MSDS turned hazardous chemicals into tradable commodities by standardizing risk disclosure about the formers’ health hazards, safe handling, storage and disposal (see UN 2019), MFSs could drastically reduce procurement friction, allowing enterprises to automate risk assessment and instantly compare vendors on a level playing field.

Finally, to complete the transition of AI from a bespoke service to a true utility, the industry must solve the "pricing fog" of inconsistent metering. Currently, providers price inconsistently: by tokens, inference time, compute type, or characters, making direct price

comparison impossible (see, e.g., OpenAI 2025g). Instead, a standardized, abstract unit of "AI work" could factor in model size and computational complexity to create a single fungible metric, which might translate into perfect substitutability for downstream firms, allowing them to switch vendors without friction—paralleling how the “kilowatt-hour” transformed heterogeneous electric power generation sources into a universally tradable commodity (see Blind 2004). Such a development would enable businesses to budget reliably and, more importantly, create a transparent, liquid market for compute where providers are forced to compete on a clear cost-per-unit basis rather than hiding behind opaque pricing models.

In terms of logistics, what might these universal data containers, standardized safety benchmarks, and fungible billing units look like in the later part of the 2020s and beyond? To operationalize them, the industry must standardize definitions of safety failures and liability assignments between model providers and enterprise users. We may therefore see the maturation of globally harmonized “AI safety contracts,” akin to the ISDA Master Agreements that standardized the global derivatives market (OECD 2024; NIST 2023).[260]

[260] ISDA stands for the International Swaps and Derivatives Association, and the ISDA Master Agreement is a standardized contract used to govern over-the-counter derivatives transactions (see ISDA 2002). Before its adoption, parties had to negotiate complex legal terms for every single trade, creating massive friction and legal uncertainty. By establishing a universal template that pre-defines critical terms—such as what constitutes a "default" or how disputes are settled—the ISDA Master Agreement transformed derivatives from bespoke, high-risk legal gambles into a liquid global market (Golden 2012).

In this scenario, SSOs would develop and governments would endorse frameworks for AI models to self-attest to certain safety and ethical parameters before deployment. We might also see the emergence of international "AI liability consortia"—modeled perhaps on the nuclear industry's insurance pools—to pre-negotiate and standardize responsibility assignments for AI failures in complex, multi-agent systems (EU 2024). This would move the industry beyond vague tort law into proactive indemnification pathways. Furthermore, to democratize access, governments could emulate the CERN model of physics by funding "public good AI foundations" (UNESCO 2021; Bengio et al. 2023). These SSO-managed bodies would curate and open-source highly capable, ethically vetted foundation models, ensuring that the "public utility" layer of intelligence remains accessible to all, rather than enclosed by a few tech giants.

An Uncertain Future Awaits AI Standardization

However, a distinct tension looms over this optimistic trajectory of seamless interoperability. While the economic logic of standardization points toward a unified global stack where agents and robots collaborate frictionlessly, the political logic of the mid-2020s points toward fragmentation. As I will discuss in the book's Conclusion, the "One World" digital order that allowed protocols like TCP/IP to conquer the globe is rapidly disintegrating into competing regulatory blocs—a world defined by the diverging philosophies of the United States, China, and the European Union.

Consequently, the final chapter will return to this theme to argue that the next phase of diffusion may require more than just technical protocols like ONNX or MCP. It may require a robust, state-led industrial strategy to navigate these geopolitical fissures. There I contend that the true winner of the AI race will not be the polity that merely creates the smartest model, but the one that successfully harmonizes these standards to win the "implementation race"—

embedding intelligence into the physical bedrock of hospitals, power grids, and factories despite a fracturing global order.

Necessary but Not Sufficient: The Role of Organizational Transformation

Putting standardization aside, the widespread availability of "utility-grade" AI revealed a deeper, more stubborn bottleneck: the organizational capability required to wield it. As of late 2025, the primary constraint on AI uptake had shifted from technical feasibility to structural inertia. Plugging an autonomous AI agent into a fragmented, pre-cloud IT stack meant that an organization's wheels could grind to a halt: While standards like OMF solved data exchange between vendors, they could not transform legacy workflow processes within enterprises.

Instead, to fully exploit a technology like AI it takes investments in "intangible capital"—redesigning workflows, retraining staff, and restructuring hierarchies (see Brynjolfsson et al. 2021). Consequently, for any single organization to realize the efficiency gains promised by AI, it had to move beyond mere "adoption" (buying the tool) to "transformation" (changing the work)—a slow, painful cultural shift that standards facilitate but cannot automate (see Appen 2021; McAfee 2019). For example, the challenge was no longer just hiring "data scientists" to build models (a 2020s problem); instead, it was hiring "AI Systems Architects" capable of redesigning specific business processes to accommodate non-human actors (Sathish and Rutten 2024).

In what follows, I examine three industries that already relied heavily on AI in 2025, manufacturing, healthcare, and finance, highlighting both how they partially automated key processes and the limitations that kept them from fully capitalizing on AI. I explore how the convergence of the technical standardization and organizational reforms discussed above may, if it indeed happens, finally allow these sectors to overcome those hurdles. Finally, I project the

transformative consequences of a more mature and democratized AI infrastructure: from near-complete factory automation to seamless supply chain orchestration, and from frictionless financial lending to hyper-personalized healthcare solutions.

Blending Standards with Organizational Reform in Manufacturing

By 2025, manufacturers were increasingly leveraging AI for quality control, especially through computer-vision–enabled automated optical inspection—systems where high-resolution cameras scan products for microscopic flaws faster than the human eye—to detect defects in electronics manufacturing (Research and Markets 2025). Firms such as Jabil piloted AI-enhanced AOI workflows to reduce false positives and improve inspection speed and accuracy, while keeping humans focused on higher-value judgment where needed (Microsoft Research 2018). In adjacent built-environment workflows, construction platforms also expanded AI use: for example, Procore launched AI capabilities to automate routine tasks and surface predictive insights across construction workflows (Procore Technologies 2024). In particular, Procore's "Insights" layer explicitly targets earlier visibility into budget performance and potential cost overruns by analyzing operational signals such as RFIs—formal requests for information used to clarify design details—submittals, and daily logs (Procore Technologies n.d.).

Continued standardization was critical to maintaining this progress, the seeds of which were already sown by 2025. With consistent labels and telemetry schemas across domain shifts, transfer learning approaches—techniques that allow a model trained on one product line to apply its "knowledge" to a new product line without starting from zero—can reuse learned defect representations across related products and settings more efficiently than training from scratch (Mih et al. 2023). Fortunately, ISA-95 (IEC 62264) provides common terminology and information exchange models for integrating enterprise planning with manufacturing operations,

reducing the need for expensive customization between IT (business planning systems) and OT (factory control systems) layers—effectively acting as a translator between the office and the factory floor (ISA 2025). Complementary interoperability approaches emphasize semantic interoperability—ensuring that data carries its own context (e.g., identifying a number as "temperature in Celsius" rather than just a raw value)—so equipment and software can exchange structured information consistently across vendors and environments (OPC Foundation 2023).

However, more intensive standardization alone is insufficient without organizational transformation to exploit it: in 2025, scaling AI in production environments still faced persistent organizational frictions. Industrial data often remained trapped across disconnected systems—creating silos, mismatched formats, and reconciliation burdens that slow or undermine analytics and AI at scale (Actian 2025). Human labor also remained essential where tasks demand fine motor skills, dexterity, and the ability to interpret ambiguous information—capabilities that continue to complement, rather than be fully replaced by, current industrial robotics and AI systems (ASME 2025). Even in inspection workflows, reducing false positives and handling borderline cases commonly requires keeping humans in the loop for review and feedback rather than expecting vision models to resolve every ambiguous example autonomously (Microsoft Research 2018). Finally, while firms are leaning into automation and AI to drive productivity, smaller companies can be constrained by financing and technical expertise, slowing adoption of more complex approaches (WCMP 2025).

The path to breaking this impasse lies in the convergence of technical standardization and organizational redesign. On the technical side, robust interfaces and shared terminology between enterprise and control layers are needed to reduce the risk, cost, and errors associated with one-off integrations—custom, fragile software bridges built for a single purpose (ISA 2025).

Organizationally, firms must reorganize around unified data foundations and decision loops that connect plant-floor signals to planning and continuous improvement cycles, rather than leaving AI tools isolated at the station level (Ugwi 2025).

Consider an auto parts supplier that executes this dual strategy. By standardizing naming conventions and data definitions across inspection stations, it becomes feasible to aggregate production and quality signals across lines and exploit these for continuous optimization. For example, when a defect is detected in a casting mold, the unified data layer allows the AI to instantly trace the error back to a specific temperature variance in the upstream furnace and flag it for maintenance—a root-cause analysis that previously required hours of manual spreadsheet reconciliation.[261] This allows the firm to shift experienced workers toward supervisory roles that deal with exceptions and validate changes, culminating in increased productivity (WCMP 2025).

Healthcare and AI

By 2025, AI had been well integrated into healthcare. Predictive AI had become mainstream in hospital operations—most notably through Electronic Health Records (EHR)-embedded risk scoring and automation, or algorithms running directly inside digital patient charts to flag potential dangers like readmission or sepsis. Adoption saw especially rapid growth in administrative and operational use cases like billing and scheduling (Chang et al. 2025). In imaging, the most visible "success" was in workflow-level integration efforts—using technical

[261] Root cause analysis is a systematic process for identifying the fundamental origin of a defect or failure. Rather than just addressing the immediate symptom (such as discarding a bad part), it seeks to correct the underlying process issue (such as a specific temperature setting or machine calibration) to prevent the problem from happening again.

blueprints to ensure AI findings appear automatically on a radiologist's screen rather than requiring a separate login (RSNA 2025).

Much of this progress was due to standardization. The steady uptake of HL7 FHIR—a universal data format that allows different medical systems to "speak" the same language—made API-based exchange a mainstream approach for sharing health information, and it increasingly shaped how applications plugged into EHR ecosystems (HL7 International 2025). In diagnostic imaging specifically, federal interoperability guidance emphasizes the importance of standardized imaging data elements and links between imaging content and the broader health record (HealthIT.gov 2025). On the imaging side, DICOM's AI guidance makes clear that AI systems should behave as good "DICOM producers"—creating files that standard hospital storage systems can actually read and file away—generating standards-compliant outputs and metadata so results can be exchanged and managed alongside imaging studies (DICOM Standards Committee 2025).

In practice, this often means AI findings are packaged as standards-compliant DICOM objects—standardized digital files containing measurements or highlighted regions—that can be stored and retrieved via DICOMweb services rather than being trapped inside a proprietary application layer—a "walled garden" software that cannot share data with other systems (DICOM Standards Committee 2025). Industry demonstrations also showed multi-vendor exchange of AI-driven annotations via standard DICOM web protocols in real-world interoperability testing environments (Gestalt Diagnostics 2025).

Despite this progress, operational governance and post-deployment oversight remained a major bottleneck. A central "grand challenge" for AI in healthcare was building systems that could reliably retrieve clinical knowledge, reason over medical questions, and communicate

answers at a quality suitable for real clinical contexts (Singhal et al. 2023). At the same time, many teams pursued workflows that could connect patient data to standardized clinical information models so that knowledge could move across systems without bespoke integration (HL7 International 2025). Qualitative evidence from healthcare AI implementers describes a "responsibility vacuum" in which maintenance and monitoring are inconsistently assigned and often under-resourced, even as participants recognize that models can degrade over time due to drift and changing clinical conditions (Owens et al. 2025). As a result, "clinical AI" often still required significant human oversight and sociotechnical workflow design, rather than being safely "set-and-forget."

A further constraint was uneven adoption across the hospital ecosystem. National survey-based reporting shows a persistent "digital divide" in predictive AI use: system-affiliated hospitals reported substantially higher adoption than independent hospitals, and rural adoption lagged behind urban facilities (Chang et al. 2025). Radiology leaders also emphasized that interoperability is often the limiting factor for realizing value from imaging AI—standards like DICOM and HL7 help, but consistent implementation guidance (via IHE profiles) is often needed to avoid incompatible vendor choices that keep AI outputs siloed (RSNA 2025).

Achieving the next stage of AI adoption in healthcare therefore requires both robust technical standards and organizational reimagining. On the technical side, aligning imaging AI outputs with standards-based integration patterns—so that AI results can flow into reporting and downstream clinical systems—remains a core interoperability objective (ibid). On the organizational side, closing the AI adoption gap will require clearer, scalable approaches to monitoring and governance—so that hospitals can assign accountability, detect degradation, and

maintain safe performance over time without relying entirely on informal, ad hoc practices (Owens et al. 2025).

Finance and AI

By 2025, AI and machine learning were widely deployed across financial services for tasks such as fraud detection, trading analytics, and customer service automation (Tierno 2024). Financial institutions increasingly combined structured data with alternative and unstructured sources—messy, text-heavy inputs like news articles, earnings call transcripts, or social media posts—where appropriate and permitted (Tierno 2024). Algorithmic and high-frequency trading relied on ultra-low-latency data processing and very short-horizon signals, identifying fleeting micro-opportunities to buy or sell securities more efficiently (Kearns and Nevmyvaka 2013). In fraud and financial-crime contexts, machine learning systems were used to detect anomalies and triage large volumes of transactional activity, alongside governance and oversight requirements (Tierno 2024).

Standardization of financial messaging significantly strengthened the data foundations for automation in payments and screening. Indeed, SWIFT's ISO 20022 migration for cross-border payments—a shift to a common global language for financial data—is explicitly framed as enabling richer, more structured data and improved interoperability (SWIFT 2025). It supports more detailed and structured remittance information, including fields that can carry invoice- and reconciliation-relevant context—meaning a payment carries the "why" and "what for" inside the message itself rather than requiring a separate email attachment (see ASC X9 2022). As richer ISO 20022 message content became more prevalent, financial institutions reported sizable reductions in compliance screening alerts attributable to more complete, structured incoming message data (SWIFT 2025).

Despite these inroads, adoption of AI aided systems remained uneven and constrained by fragmented data architectures and governance requirements (Tierno 2024). A persistent challenge was explainability and documentation for “high-stakes” and customer-facing decisions, where supervisors emphasize model risk management expectations (Pérez-Cruz et al. 2025). Consider that in U.S. credit decisions, adverse action notices must provide specific and accurate principal reasons for denials even when complex or AI-driven models are used, which creates practical pressure for traceability and defensibility—forcing banks to prove exactly why the computer said "no" (CFPB 2023). The practical implication is that a financial institution’s AI model choices and explanation practices are shaped by transparency and audit expectations (Pérez-Cruz et al. 2025).

Looking forward, there were still opportunities for further standardization and organizational adaptations in the financial services industry. Semantic standards and "policy/compliance-as-code" practices represent plausible pathways for further progress. The so-called Financial Industry Business Ontology is designed to define shared financial concepts and relationships in a machine-readable way to support semantic interoperability—functioning essentially as a digital dictionary that ensures a term like "interest rate swap" is defined identically for a computer in London as it is for one in New York, preventing costly translation errors (EDM Council, n.d.). Compliance-as-code (often discussed as "policy as code") refers to expressing controls and requirements as machine-readable rules that can be enforced and evidenced through automated workflows—essentially translating dense legal rulebooks into software code so that regulatory violations are caught and flagged instantly by the system rather than by a manual audit months later (Contino 2021). In regulated contexts, these approaches still require human oversight for interpretation, exceptions, and governance (Pérez-Cruz et al.

2025)—creating opportunities to reform financial institutions to be more amenable to AI aided automation.

Envision a mid-tier regional bank that unlocks this potential. Historically dependent on lagging, batch-oriented underwriting signals—credit scores updated only once a month—the bank modernizes its payments and data backbone by adopting ISO 20022–native infrastructure to preserve richer, structured payment context end-to-end (SWIFT, 2025). With standardized, data-rich message fields flowing cleanly through its systems—including structured remittance information—the bank can more reliably automate downstream analytics and controls (ASC X9, 2022).

That upgraded data layer makes it feasible to participate in a privacy-preserving, cross-institution fraud-learning setup using federated learning—a technique where algorithms learn from data on local servers without the sensitive customer records ever leaving the bank's secure vault—collaborating on model improvements without pooling raw customer data (Aljunaid et al. 2025). Now, when a customer applies for a loan, the bank's models can incorporate transaction-derived cash-flow features alongside traditional credit bureau variables, which empirical work finds can improve predictiveness relative to bureau-only approaches (FinRegLab 2025).

The bank operationalizes explainability and documentation as part of model risk management—because supervisors emphasize that limited explainability in complex AI can create governance and compliance challenges (Pérez-Cruz et al., 2025). For adverse credit decisions, it generates specific, accurate principal reasons that can be communicated in plain language to applicants, consistent with rigorous U.S. regulations even when complex models are used (CFPB 2023). The result is a faster, more data-informed decision process that can reduce

friction for creditworthy applicants while maintaining auditability—narrowing the experience gap that has often favored digital-first lenders (Tierno 2024).

CONCLUSION

By 2025, the "AI revolution" was no longer a speculative narrative; it was a physical reality cast in silicon, steel, and copper. The unprecedented capital wave documented in this chapter successfully industrialized the production of intelligence, driving marginal costs down and accessibility up. Through the hard work of standardization—from the atomic level of data formats to the complex orchestration of autonomous agents—the economy built the "rails" upon which AI could travel. We witnessed the start of a transition from bespoke, fragile models to robust, plug-and-play utilities, and we saw the early green shoots of transformation in sectors as diverse as high-frequency trading and high-volume manufacturing.

Yet, the successful construction of this machine revealed a new set of frictions. As the "Standard Setters" (model developers seeking frictionless access) clashed with the "Gatekeepers" (incumbents seeking to capture value), it became clear that the standardization of AI is as much a political project as a technical one. Moreover, building the machine is not the same as benefiting from it. As the sectoral analyses of healthcare, finance, and manufacturing demonstrated, the existence of capable technology is a necessary, but not sufficient, condition for economic revolution. The primary constraint has shifted from the technological (can the model do it?) to the organizational (can the firm adapt to it?).

With the physical infrastructure built and the marginal cost of intelligence approaching zero, the barriers to automation are rapidly dissolving. This leads us directly to the most pressing questions of our time, which I address in the next chapter: As this newly standardized, industrial-scale intelligence permeates the labor market, what happens to the value of human work? Will

this new abundance lead to shared prosperity, or will the "Gatekeepers" of the new infrastructure capture the surplus, supercharging inequality?

CHAPTER 11. MACROECONOMICS OF AI

If you had looked at many of the headline numbers I've documented across this book, everything AI related took off in a steep ascent during the early 2020s: investment dollars behave exponentially, as do chip speeds, model capabilities, price declines, and the uptake of vanguard models.

Yet, notwithstanding the productivity green shoots I document later in this chapter, when we look at the actual economic scoreboard, the results are surprisingly muted—echoing Solow's eponymous Paradox (Solow 1988; Chapter 2 of this book): the AI boom was everywhere, except in the productivity statistics.

To be sure, as Chapter 10 made clear, in 2025 the economy was undergoing a massive bout of what economists refer to as "capital deepening." Companies were spending billions on GPUs, data centers, and software licenses. In the national accounts, this looks like an explosion of inputs—throwing more money and machinery at AI to accelerate its technical progress.

But spending money on a supercomputer does not automatically make a business more efficient; it just makes it more capital-intensive (see Brynjolfsson et al. 2021). And though US firms experienced up to 40 percent productivity gains in knowledge work after automating tasks like software coding and data analysis (Brynjolfsson et al. 2025), only 25 percent of US tech jobs listed in early 2025 were seeking employees with AI skills (Radauskas 2025). To be frank: for many, AI seemed like a productivity bust.[262]

[262] But not to everyone! Workers using generative AI experienced an average time savings of 5.4 percent of work hours (Bick et al. 2025).

The solution to this puzzle lies in the shape of the curve we are riding. Those experiencing the evolution of General Purpose Technologies often mistake the early, vertical burst of a technology for a permanent trajectory. But history suggests that an S-curve is the best fit to describe its ultimate path (see Griliches 1957). In the beginning, growth *feels* exponential because we are picking the "low-hanging fruit": early adopters who are tech-savvy and use cases that are easy to automate dominate both the discussion and data. But eventually, the easy wins run out. The adoption curve inevitably flattens, not because the technology stops working, but because we hit the hard part: highly regulated industries, messy legacy data systems, and organizations that are slow to change.

As I will demonstrate in the formal model at the heart of this chapter, this transition from "easy adoption" to "hard standardization" is exactly what the math predicts, matching the historical evolution of GPTs outlined in Chapter 2 and suggesting that a more muted, mundane AI future awaits.

True improvements in productivity, the kind that significantly lift wages and living standards, only happen when that new capital is paired with a new way of working (see Bresnahan et al. 2002). This is the difference between the *Frontier* (what the best AI model can do in a lab) and *Diffusion* (what the average worker actually does with it on a Tuesday morning), a topic I explored in the previous chapter.

Will There be Jobs?

Supposing that AI does usher in a productivity renaissance, whither the jobs? Several leading C Suite Executives warned that AI-driven productivity gains would reduce future head counts. Amazon CEO Andy Jassy stated that the company expected its corporate workforce to shrink over the next few years as it adopted generative AI tools and agents (Palmer 2025). Similarly,

JPMorgan Chase's Marianne Lake told investors that, due to AI, the bank's workforce could shrink by 10% (Cutter and Zimmerman 2025). Most dramatically, Anthropic's CEO, Dario Amodei, warned that AI could eliminate up to 50% of entry-level white-collar jobs within five years and push unemployment to 10–20% (Martin 2025). Ford CEO Jim Farley made similar claims (Cutter and Zimmerman 2025).

When these influential business leaders made these dire forecasts, they had 2025's increasingly advanced AI models in mind. Because these models excel at language processing, data pattern recognition, and standardized reasoning processes, they have the potential to disproportionately affect higher-educated, higher-paid knowledge workers. These include financial analysts, underwriters, compliance officers, market research analysts, sales representatives, and even content creators (Cazzaniga et al. 2024; Frey and Osborne 2013).

CEOs' warnings about the disruption that awaited white-collar professional workers appeared to be grounded in a growing body of evidence (Hoffman 2025; Leopold 2025). As early as 2022, AI task automation in everything from voice-recognition and speech-to-text algorithms to automated guided vehicles in logistics had reduced employment in both low-skilled and middle-skilled occupations, as well as lowering workers' wages in these categories (Marguerit 2025).

Over the following years, entry-level roles were especially affected. On Handshake, entry-level job postings fell about 15% year over year while applications per posting rose about 30% between the 2023–24 and 2024–25 school years, signaling fewer opportunities and tougher competition (Handshake 2025). Indeed, Brynjolfsson et al. (2025) confirm this pattern with hard data: since the proliferation of generative AI, early-career workers (ages 22-25) in the most AI-exposed occupations like software development and customer service have experienced a 13%

relative decline in employment, a trend robust to firm-level shocks and not explained by broader economic factors. Consequently, during 2024-25, recent college graduates' unemployment rate was running significantly higher than the overall workforce, at about 6% versus just above 4%, a notable reversal relative to typical historical patterns (Peck 2025).

These trends were consistent with increasing AI adoption in white-collar work. There were fewer employment opportunities for entry level workers in technical fields such as finance and computer science, where AI adoption was most rapid (Leopold 2025). At Big Tech firms, new-grad hires represented about 7% of total hires in 2024, with entry-level hiring down roughly 25% versus 2023, and more than 50% below pre-pandemic (2019) levels; at startups, new-grad hires were under 6%, down about 11% year-over-year and over 30% since 2019 (Bantock 2025). The same pattern also affected other industries such as financial services and consulting where "grunt work"—social-media copywriting, basic coding, and routine report generation—was increasingly handled by AI tools instead of new hires (Hoffman 2025; Martin 2025).

The upshot was stark: if AI could automate many of the tasks that once served as training grounds for junior employees, then entry-level roles that had traditionally provided on-the-job training and mentorship might disappear, along with pathways for upward mobility (Hoffman 2025). Raman (2025) argued that this phenomenon risked long-term earnings losses and greater inequality, likening it to the 1980s collapse of US manufacturing when factory automation combined with globalization to induce a steep decline in blue-collar entry-level work. He lamented that the loss of those once-reliable first-step positions not only depressed wages for a generation of young workers but also triggered social and political upheavals.

Indeed, there were even starker warnings issued by some analysts such as Mac Ghlionn (2025). He warned that while in 2025 skilled tradespeople such as electricians were shielded

from AI because robots still struggled with tasks that required fine motor skills, dexterity and the ability to troubleshoot on the fly in messy, real-world environments, as embodied AI and robotics improved, many blue-collar jobs could potentially be automated. Mac Ghlionn imagined AIs with improved vision, strength and torque power such as robot plumbers that sense pressure changes in pipes and adjust torque mid-job; robot welders that detect imperfections and adapt parameters on the fly; and drones laying shingles with laser-level precision for roofing and field work, supervised from the ground. The upshot would be much fewer entry-level rungs for trained apprentices and the eventual demise of craftwork for humans.

This glum perspective mirrored public sentiment. A 2025 Pew Research Center survey found that 64% of U.S. adults believed AI would lead to fewer jobs in the next 20 years, while only 5% expected more of them (McClain et al. 2025).[263] Overall, 56% of U.S. adults expressed extreme or very high concern about AI eliminating jobs, indicating widespread anxiety about the technology's potential to disrupt the labor market.[264]

History's Lessons

[263] The remaining respondents were divided between those who believed AI would not make much difference (14%) and those who were unsure (16%).

[264] When examining specific occupations, public expectations of job losses varied considerably. The highest concern was for cashiers, with 73% of U.S. adults predicting fewer jobs due to AI, followed by factory workers at 67%, and journalists at 59%. As for teachers, 43% of the public expected job reductions.

At first glance, fears about AI destroying human jobs seem well founded. History is replete with examples where technological shifts rendered once-ubiquitous professions obsolete, leaving workers to face a harsh transition as their skills were devalued overnight.

The horse-drawn transportation ecosystem that dominated the early 1900s provides a striking example. Horse breeders, stablehands, blacksmiths, carriage makers, and street cleaners who removed horse manure from city streets all saw their livelihoods disappear with the advent of internal combustion engines and automobiles; the decline of horse-centered industries coincided with the rise of the automotive economy and its associated occupations (McShane and Tarr 2007).

Telephone operators experienced a similar transformation as switching technologies mechanized call routing and displaced large-scale manual switchboard work, while other clerical and service roles expanded (Feigenbaum and Gross 2022).

The role of human computers—teams of people who performed complex calculations by hand—became obsolete as electronic computers like ENIAC scaled computation and helped redirect skilled labor into early programming and related computing roles (Joosse 2022).

Agricultural fieldworkers experienced perhaps the longest and most dramatic transformation, spanning from the 1930s to the present. Mechanization increased output per worker and reduced labor needs, helping push agricultural employment down to a small share of overall employment in the modern economy (Dumont 2024).

Perhaps more apposite, consider examples from the digital age. Travel agents saw their industry disrupted from the 1990s through the 2010s as online platforms enabled travelers to research and book trips directly, limiting demand for traditional intermediary services. Similarly,

motion-picture projectionists saw their profession contract sharply as cinema projection digitized and the occupational footprint became very small.

Finally, consider the robotics explosion of the 2010s. By 2019, the global operational stock of industrial robots had swelled to 2.7 million (IFR 2020). These machines did not just weld cars; they scrubbed plates. In commercial kitchens, robotic arms paired with machine vision had begun to replace manual labor (Dickinson 2023).

Fortunately, as Chapter 2 clarified, the history of automation is much more nuanced than this portrait of "jobsmaggedon." While automation makes specific *tasks* obsolete, it often acts as a complement to *labor* by generating new, higher-value roles, often in entirely new industries. In the 19th-century textile industry, the arrival of the Power Loom did not empty the factories; instead, as the cost of cloth plummeted and demand soared, mills hired more workers to maintain the machinery and supervise complex operations—roles that demanded the acquisition of new "procedural" and "technical" knowledge. A similar dynamic played out in the automotive sector, where both the moving assembly line (1913) and the introduction of robotic arms (1970s) coincided with workforce expansions, as automakers recruited armies of industrial engineers, quality managers, and technicians to support mass-market production. Even in the service sector, the deployment of ATMs famously lowered the cost of operating bank branches, encouraging banks to open more locations and hire more tellers—transforming them from cash counters into relationship managers.

Systematically Addressing Economic Puzzles About AI

To help integrate the type of historical lessons about technological progress and jobs outlined above, as well as its impact on productivity, real wages, inequality, and profits, this chapter develops a formal framework to move beyond broad generalizations about AI; instead, I

rigorously examine the mechanics of a transition to an AI economy. Specifically, the modelling exercise I conduct ahead allows me to gain purchase on six fundamental issues that will define the ultimate size of the pie and its distribution during the Fourth Industrial Revolution.

The Three-Way Split: Labor, Capital, and Profits

First, how will the benefits of AI be split between Labor, Capital, and Profits? The model challenges the traditional binary view of the economy as a struggle solely between "Labor" and "Capital." Following the insights of Barkai (2020) and De Loecker et al. (2020), I decompose economic output into three distinct claims: the wages paid to workers, the returns paid to the owners of physical capital (servers, energy grids, chips), and a residual slice known as the Profit Share (or "pure rents").

This distinction is vital. Contrary to conventional wisdom, in the AI era the primary tension may not necessarily split workers and machines but instead divide the productive economy (both labor and capital) and the market power of dominant firms. The model demonstrates that if barriers to entry remain high, the "Profit Share" for firms that fully integrate AI can expand at the expense of both wages and capital investment, siphoning off the gains of innovation before they reach the broader economy.

The Decoupling Mechanism: Why Wages Might Lag Productivity

Second, the model mathematically formalizes the link—and the potential wedge—between what workers produce and what they earn. It isolates the conditions under which a potential "AI Productivity Miracle" translates into rising living standards versus those where wages stagnate.

If dominant firms use their market power to keep prices high despite falling costs (a "Vampire" dynamic), the model shows that real wages will mathematically flatline even as worker efficiency skyrockets. This explains the "Decoupling" phenomenon: the paradox of an

economy that is technically soaring but distributionally stagnant. Conversely, if competition forces markups down, the model predicts a "Scale Effect" where falling prices unlock massive demand, reconnecting pay to productivity.

The Engine of Prosperity: Diffusion vs. Friction

Third, the model identifies technological diffusion as the critical lever for economic growth during the AI revolution. Here, it takes a page from Chapter 2, where history taught us that technologies only revolutionize economies when they become standardized, cheap, and ubiquitous. The internet and mobile computing changed the world not just because they were useful, but because prices fell relentlessly, fueling massive adoption. This stands in contrast to a "Dual Economy" scenario—where diffusion stalls, driving the frontier apart from the fringe.

To gain purchase on this idea, consider Japan's enduring economic stagnation or "Lost Decades" again, which I discussed in Chapter 4. On one track, export manufacturers like Toyota and Sony were able to achieve phenomenal productivity levels that rivaled or beat the global frontier thanks to fierce competition (Porter and Sakakibara 2004). On the other track, the domestic service sector (retail, construction, and banking) lagged badly. Shielded from competition by regulations like the Large-Scale Retail Store Law (*Daiten-ho*), these firms failed to adopt the IT revolution that drove U.S. growth (Katz 1998; McKinsey Global Institute 2000). Because the knowledge and efficiency of the export giants never diffused to the domestic fringe, the economy bifurcated.

My model captures this peril by formalizing the tension between Standardization (which accelerates diffusion) and Friction (which creates scarcity). It proves that the ultimate impact of AI depends on a specific race: Can the rate of diffusion accelerate faster than the rate at which "Gatekeepers" entrench their market power? If not, the AI economy may usher in

“Japanification”: a permanent state of high-tech enclaves surrounded by low-productivity stagnation.

The Alignment between Profit Creation and Prosperity

Fourth, the model illustrates the "Gatekeeper's Dilemma." By keeping AI expensive or proprietary in a bid to amplify their markups, dominant firms strangle the diffusion process required to grow the market. This in turn heralds a "high-margin, low-adoption" equilibrium where the technology remains a luxury good rather than a general-purpose utility.

Indeed, history is littered with the carcasses of technically superior inventions—from the Segway, LaserDisc, and Iridium satellite phones to the Concorde and high-end VR headsets—that suffocated in this “boutique trap,” remaining expensive curiosities for the wealthy rather than transforming the daily lives of the many.

Therefore, a "Cyber Robber Baron" scenario—where a few firms get rich by keeping AI expensive and exclusive—is not just bad for labor; it is bad business. The profit-maximizing move for the ruthless capitalist is, paradoxically, to democratize the technology as fast as possible. A true superstar firm realizes that in a highly elastic AI market, the only way to maximize the size of their slice is to relentlessly expand the size of the pie—mirroring the strategy that drove the PC and smartphone revolutions documented in Chapters 5 and 6.

AI and Jobs: The Race Between Substitution and Scale

Fifth, the framework tackles the anxiety surrounding AI and Employment. It moves beyond the simplistic "robots take jobs" narrative to model the tug-of-war between the Substitution Effect (displacement) and the Scale Effect (expansion). The model predicts that if "Gatekeepers" hoard efficiency gains, we get displacement without expansion. But if diffusion lowers prices, the resulting demand surge creates a complementary role for human labor based on Comparative

Advantage. In this scenario, the economy does not simply need fewer workers to do the same thing; it uses the efficiency dividend to do more things, revaluing human judgment and oversight in the process.

Reconciling the "Superstar" and the "Gatekeeper"

Finally, the model's focus on market power requires a nuanced distinction. Readers will recall that in Chapters 4 and 5 I championed the "Superstar Firms" of the Third Industrial Revolution—companies like Intel and Qualcomm—as engines of prosperity. I argued that their high profit margins were rewards for superior efficiency in a competitive market that, through mechanisms like patent citation networks and Standards Setting Organizations (SSOs), eventually diffused innovations to the rest of the economy.

The formal model presented here tests the limits of that dynamic. It distinguishes between the "Virtuous Cycle" of the Superstar (where high profits fund the diffusion that expands the pie) and the "Vicious Cycle" of the Gatekeeper (where high profits rely on blocking diffusion to capture a larger slice). This insight moves beyond formless optimism or pessimism, instead offering a precise set of conditions that ignite a high-growth equilibrium that lifts labor, capital, and profits alike. Specifically, if the AI ecosystem can solve its collective action problem and coordinate to lower markups by jointly assembling a standardized, interoperable infrastructure layer that gives birth to an AI agent economy, much like the 4G/LTE standards of the smartphone era fostered the technology's mass adoption and birthed the gig economy during the twilight of the Third Industrial Revolution. In this scenario, firms take a smaller slice of each transaction, but the total volume of activity explodes so significantly that their total profits rise.

This chapter continues as follows. First, I ground the analysis in the historical reality of past Industrial Revolutions, distinguishing between the "Potential Yield" of a new technology

and its "Realized Yield" after accounting for the friction of standardization. I then formalize this intuition by constructing a macroeconomic production function that decomposes AI's impact into two distinct forces: the frontier capability and the diffusion rate. Expanding this framework to distribution, I introduce the "Markup Wedge" to model the "Gatekeeper's Dilemma"—the tension between high-margin/low-adoption strategies and the high-volume "bonanza" required for broad prosperity. I then bridge this macroeconomic view with a microeconomic analysis of the labor market, using a supply-and-demand framework to demonstrate how the "Scale Effect"—and the persistence of human comparative advantage—can offset the forces of substitution, provided that market power is kept in check by competitive pressures, standardization, or a mix of both. Next, I apply this theoretical lens to adjudicate the mixed and tentative empirical evidence regarding AI's impact on jobs, separating signal from noise. Finally, I outline the limitations of this modeling strategy and propose extensions for future research.

AI AS AN INDUSTRIAL REVOLUTION

To explore the potential economic impacts of AI over the long run, I return to the historical lessons from Chapter 2—but this time around, prospectively: I employ tools from macro- and micro- economics to speculate about how AI may weave itself into the fabric of the greater economy beyond 2025. Like with steam, electricity, and computers during the first, second, and third industrial revolutions, it is likely that to better exploit AI, entire industries will rewire the way they operate and transform their supply chains, their use of labor, and business models. As more organizations embrace plug-and-play AI offerings, providers should reach economies of scale and further reduce the price of more advanced features—unlocking more value for mid-sized and small enterprises in the process that drives further adoption. This is a virtuous circle

that has characterized previous General Purpose Technologies (Bresnahan and Trajtenberg 1995).

Increased AI functionality, affordability, and mass uptake will fully materialize only if two things happen simultaneously, however: products and services become standardized, making them easier to use and worth integrating into workflows and lifestyles, and organizations adjust their culture, practices, and human resources to better accommodate, if not embrace, AI. In the cases of steam power, electricity, and computers, initial capital investments and continued performance improvements were only followed by widespread adoption of these technologies once universal technical, reliability, and safety standards were established, workers acquired complementary skills, and user-friendly, plug-and-play applications emerged.

Even so, vintage technologies' uptake trajectories are instructive for another reason too: they were relatively protracted, uneven, and surprising along the way. For example, during the Third Industrial Revolution, the progression from specialized mainframes to personal computers, then to the internet with on-site servers, and finally to cloud computing, was circuitous and unplanned. The establishment of standardized protocols such as TCP/IP—the suite of communication protocols used to interconnect network devices on the internet—was a long and winding road; the impressive economies of scale enabled by the unfurling of an extensive fiber-optic and server network only followed a speculative spending boom during the dot-com gold rush, culminating years down the road, but only in hindsight, in the cloud computing revolution. Nonetheless, the history of GPTs still offers lessons that can be systematically integrated into a theoretical framework with not only explanatory but predictive power—that is, provided it's the right framework!

Why Automation Creates More Jobs than it Destroys

As outlined in previous chapters, a GPT is a foundational innovation that transforms multiple economic sectors, improves continuously, and stimulates complementary innovations. GPTs are characterized by widespread applicability across industries, potential for ongoing refinement, and the ability to spark additional innovations in downstream markets.

There are several reasons why new GPTs typically expand rather than contract overall employment opportunities.[265] Technology fundamentally reshapes the production landscape by enlarging addressable markets through cost and quality improvements, generating secondary demand through higher incomes, creating new complementary work alongside automated systems, and optimizing task allocation based on comparative rather than absolute advantage. The net effect shifts the nature and location of human work rather than reducing the overall need for human economic participation. Let us examine each of these reasons in turn.

First, consider demand elasticity. When automation reduces unit costs, firms naturally lower their prices to consumers. In markets where demand responds elastically to price changes, the resulting increase in quantity demanded often exceeds the proportional price reduction, leading to expanded output that requires additional human labor upstream in supply chains, downstream in sales and service operations, and alongside automated systems to manage the larger market volume.

Second, consider higher effective demand. Beyond simple cost reductions, technological advances frequently enhance product quality by making offerings more reliable, personalized, or feature-rich. These improvements shift consumer demand curves outward as people become

[265] For a canonical review of these compensation mechanisms—and the historical evidence that automation has generally complemented rather than substituted for labor—see Autor (2015).

willing to pay more for superior products, creating higher effective demand that supports expanded production and generates additional human jobs in design, customization, delivery, and customer support for these enhanced offerings.

Third, consider that the productivity gains from human-technology collaboration creates a powerful multiplier effect throughout the economy. As workers leverage new tools to increase their output per hour, competitive labor markets translate a significant portion of these productivity improvements into higher wages. This additional disposable income flows into other sectors—restaurants, travel, healthcare, education—creating employment opportunities in industries that expand specifically because of the income effects generated by technological advancement.

Consider a simple example. Suppose a grocery chain implements AI-powered self-checkout kiosks that handle half of all register transactions. Labor costs for cashiers drop, allowing the chain to reduce prices on groceries. Shoppers then use their savings to dine out more, join fitness classes, or hire home services—all of which require human workers. Thus, automation in grocery checkouts results in lower food prices, higher real incomes, and increased spending on other sectors, creating new jobs in restaurants, gyms, and home-care services. In short, productivity gains due to automation lead to lower prices, higher wages, and increased consumption, which in turn create new jobs (Atkinson 2025; Kagan 2024; Wolla 2020).

Fourth, consider that even in highly automated environments, human workers remain essential for what might be called "last mile" activities that amplify machine capabilities. Workers coordinate complex production processes, diagnose unusual situations that fall outside programmed parameters, ensure quality standards, manage customer relationships, navigate regulatory compliance, maintain physical facilities, and continuously optimize equipment and

production processes. These complementary roles may grow in importance as systems become more sophisticated and operate at greater scale and complexity.

Take textile manufacturing during early American industrialization. This industry showcases how partial automation paradoxically rewarded human expertise. Bessen (2015) shows that while technological innovations automated certain aspects of bobbin winding and shuttle loading on looms, it simultaneously elevated manual tasks like tying particular knots, optimizing manual operations through repeated practice to efficiently replace empty shuttles, and acquiring technical knowledge through experimentation—learning to adjust warp tension to minimize thread breakage and coordinate work across multiple looms. Not only did these weavers command substantially higher wages than their pre-automation counterparts, but the surge in textile demand due to the higher quality products associated with the combination of upskilled labor and automation caused mills to hire many more of them.

Fifth, consider that the principle of comparative advantage ensures that optimal resource allocation doesn't simply assign all tasks to whichever performer has absolute superiority. Since machine capacity, energy, and computational attention remain scarce resources, deploying a system for one task carries the opportunity cost of foregone value creation elsewhere. Economic efficiency requires dividing work so that machines handle activities where their relative advantage is greatest, while humans focus on tasks where reassigning machine resources would sacrifice higher-value applications. This division of labor maximizes total economic output while sustaining robust demand for human workers.

HOW AI MAY IMPACT THE PRODUCTIVITY OF THE ECONOMY

U.S. productivity growth has experienced long, discordant regimes during the post World War II era: rapid postwar ascent, a slowdown after the early 1970s, an IT-era acceleration in the late

1990s, and then a marked downshift after the mid-2000s (Gordon 2016). What distinguishes the post-2005 period is that the underlying efficiency term—total factor productivity (TFP)—barely advanced: cumulative TFP rose only about 5% after 2005, compared with roughly 55% growth from 1955 to 2015 (see Gordon 2016; Federal Reserve Bank of St. Louis 2024).[266] Indeed, private nonfarm business TFP increased by 0.7% in 2023, roughly in line with its pre-pandemic 2011–2019 pace, rather than a return to the late-1990s boom (BLS 2024).

Could AI reinvigorate productivity, an imperative in the face of aging populations, tightening labor supply, and a less integrated global economy? Perhaps. Approaching 2025, nonfarm business labor productivity rose 2.3% in 2024 (annual average), and increased at a 3.3% annualized rate in the second quarter of 2025 (BLS 2025b; BLS 2025c).[267] While it was too early

---

[266] The BLS calculates total factor productivity (TFP) using a growth-accounting approach that compares how fast output grows to how fast the inputs used to produce it grow (BLS 2011). Specifically, TFP growth equals output growth minus the weighted sum of input growth rates, where each input is weighted by its share of total costs. Labor costs include wages, salaries, and benefits; capital costs include depreciation and the cost of using equipment and structures; and, in some BLS industry measures, purchased inputs like energy, materials, and outside services are also counted (ibid). This calculation isolates the portion of output growth that cannot be explained by measured inputs—a residual that reflects improvements in efficiency, technology, and related factors (ibid).

[267] The BLS defines labor productivity (output per hour) as real output divided by hours worked. For the nonfarm business sector, the output index is based on BEA's real value-added business-sector output derived from GDP after excluding general government, households, and nonprofit

to attribute these gains to any single cause by the time of this writing, the timing renewed interest in whether AI adoption is beginning to lift measured productivity (see Cook 2024).

The strongest evidence for a role played by AI came from micro settings where AI tools are deployed and outcomes can be measured. In a large field study of customer-support work, access to a generative-AI assistant increased productivity by about 15% on average, with disproportionately larger gains for less-experienced workers (Brynjolfsson et al. 2025). Survey evidence was also consistent with meaningful time savings: nationally representative data suggest that workers report time savings equivalent to roughly 1.4% of total work hours from generative-AI use (Bick et al. 2025). In sectoral productivity accounts, some service industries—including health care and social assistance—also recorded positive total factor productivity growth in 2023 (BLS 2024), though linking those gains specifically to AI remains an open empirical question.

Scaffolding an AI-Centered Macroeconomic Model

When Artificial Intelligence (AI)—or any GPT—enters the scene, it has the potential to raise productivity across multiple sectors. Consider that labor productivity is driven by technological changes, capital investment, capacity utilization, use of intermediate inputs, improved managerial skills, and enhanced workforce capabilities (Syverson 2011). Total Factor Productivity (TFP) growth reflects improvements in the efficiency with which all these inputs are combined to produce output, serving as a measure of technological progress, organizational

---

institutions serving households; BLS then excludes the farm sector to obtain the nonfarm business sector (BLS 2025d; BLS 2025e).

improvements, and other factors that enhance overall efficiency. In turn, this depends on R&D, capital expenditures, and education and training (Acemoglu 2009).

AI parallels earlier GPTs like steam power, electricity, and computing because it can transform how labor and capital combine to produce goods and services, creating new opportunities for automation and augmentation. Automation captures AI's ability to substitute for labor in routine or data-intensive tasks. Augmentation occurs when AI tools help humans do higher-level tasks better—e.g., faster R&D, deeper analytics, or new product design (Filippucci et al. 2024).

However, previous industrial revolutions centered on earlier GPTs remind us that it's also important to consider the political economy of technological diffusion. Broad-based productivity gains typically lag huge bursts of capital spending. Indeed, productivity payoffs from a major innovation might take years (or longer) to show up in the aggregate data—a phenomenon often dubbed the "productivity paradox" (Crafts 2021). In the long run, once a General Purpose Technology matures after standardization and organizational adjustments allow it to fully diffuse, it can lift TFP growth by reorganizing how work is done and expanding what is possible across the entire economy.

Therefore, below I introduce a framework that differentiates between the short- and long-run, mirroring an S-curve adoption dynamic that is typical of GPTs (Lipsey et al. 2005). Early on, uptake is limited to well-resourced organizations that can absorb high upfront costs and navigate unproven technology. The lion's share of organizations will only adopt and fully exploit a GPT belatedly, however, as they await the technology to become standardized and they make complementary investments and overhaul their organizations—adopt new workflows, retrain staff, and obtain buy-in from managers and employees—and supply chains. Adoption grows

faster when more of the economy has already adopted as these practices create demonstration effects, opportunities for learning-by-doing, imitation, and as “best practices” spread (Arrow 1962; Rogers 1962). Finally, adoption must inevitably slow down as the economy approaches full adoption, because there are fewer non-adopters left, and because the remaining holdouts are often the hardest cases: they must rely on legacy systems due to the demographics of their workers or users, are subject to heavy regulation, or face credit market imperfections that make it cost prohibitive for them to upgrade their technology.

It is important to clarify the theoretical lineage of this approach. My contribution is not a novel growth theory, but an application of diffusion mechanics to the specific institutional frictions that currently constrain the AI economy around standardization and organizational redesign. On the one hand, I retain the Solow framework’s production structure by incorporating a measured, economy-wide effective AI-enhanced “technology term” that scales output for given inputs of labor. On the other hand, I depart from its assumption that technology falls like "manna from heaven," instantly available to all (see Solow 1957; Acemoglu 2009), by building directly on the "technology diffusion" literature established by Nelson and Phelps (1966) and formalized empirically by Comin and Hobijn (2010).

Just as these authors model productivity as a function of the "distance to the frontier" determined by absorptive capacity or variable adoption lags, I model the realized effective technology as the frontier discounted by a “standardization wedge”. Specifically, embedded in this wedge are the lack of standards or “proto-standards” that prevent systems from talking to one another; fragmented data architectures that trap information in legacy silos ("missing plumbing"); the absence of industrial-grade pipelines to monitor models ("MLOps"); and the sheer difficulty of retooling workflows to accommodate non-human agents.

Formalizing AI Productivity, a Benchmark

Before considering the effects of AI as a GPT on productivity as an uneven and lagged process, I first consider a simpler model: a neo-classical production function where technology is exogeneous (Solow 1957). In this model, if AI raises "effective technology" it mechanically raises the productivity of both capital and labor—not by changing how inputs combine, but by making the whole input bundle produce more. This is the standard growth-accounting interpretation of technology as a multiplicative shifter of output. Labor productivity can rise because: AI makes the economy more effective at turning inputs into output and/or workers have more capital to work with. And if AI is embodied in higher effective technology, then this should translate into higher TFP. Moreover, if you have a lot of capital and labor "in play," then improving the technology that organizes these amounts has a bigger impact on total output than if the economy possessed smaller stocks of capital and labor to leverage. This baseline model therefore establishes a clean starting point before I introduce diffusion/standardization frictions later: I establish what would happen if AI works perfectly everywhere (the "Potential Yield"); then add "wedges" that explain why we may not be seeing that in the data yet (the "Realized Yield").

Following convention (see Solow 1957; Acemoglu 2009; Syverson 2011), I can formalize this idea as:

$$Y_t = A_t^{eff} F(K_t, L_t) \quad (1)$$

Where:

- Y_t is aggregate output at time t.
- K_t is capital input (equipment, structures, software, and—depending on national accounts—some intangible capital).

- L_t is labor input (hours worked, or efficiency-adjusted labor if you later want to incorporate human capital).
- $F(\cdot)$ is a generic production function describing how capital and labor combine to produce output.
- A_t^{eff} is an **effective technology term** capturing *how productively the economy can turn inputs into output*.

In this baseline model, when I say that AI raises A_t^{eff}, I am assuming frontier AI capability is standardized, diffused, and embedded across firms; later, I will relax this assumption, varying how much of that technology's capability is actually embedded in production given the degree of AI standardization and organizational assimilation costs and dynamics. Specifically, further ahead, I will decompose A_t^{eff} into a frontier component ("what's technically possible"); and a diffusion/standardization component operationalized as an adoption wedge λ_t ("what's actually happening").

For now, allow me to unpack A_t^{eff} some more in terms of the conventional growth accounting exercise: this baseline model assumes "technology is a multiplicative shifter": it scales what the economy can produce from a given bundle of inputs. That means that **AI can raise output without immediately changing measured inputs.** If A_t^{eff} rises (because AI makes workflows faster/better), output rises even if measured K_t and L_t don't. In other words, this term is the conceptual slot for TFP, the portion of output (residual) not explained by measured inputs.

Cobb Douglas Production Function

For now, consider the famous Cobb Douglas Production Function with constant returns to scale:

$$Y_t = A_t^{eff} K_t^{\alpha} L_t^{1-\alpha} \qquad \text{with } 0 < \alpha < 1 \qquad (2)$$

This function allows me to make a simplifying assumption about the form of $F(K_t, L_t)$. Specifically, the Cobb–Douglas approach is:

- **Tractable** (it provides closed-form marginal products and clean productivity decompositions).
- **Standard** in macroeconomics and growth accounting (Cobb and Douglas 1928; Solow 1957; Acemoglu 2009).
- **Interpretable:** under competitive factor markets, α corresponds to capital's income share and $1 - \alpha$ to labor's income share.
- **Assumes that technology is neutral**, which means that improvements in Total Factor Productivity raise the marginal products of both capital and labor proportionally, leaving relative factor shares unchanged.

None of this means that the real world is literally Cobb–Douglas in every sector, nor that factor shares always remain constant. Rather, Eq. (2) is a **benchmark** that will let me cleanly show what "terms" I must add to the model to generate richer distributional outcomes; specifically, later ahead I will modify the model with terms that represent "biased technical change," "market power," and varying "task structures," etc.

With $Eq.(2)$ in hand, I can now compute:

- the marginal product of capital (MPK),
- the marginal product of labor (MPL),
- and then show how changes in A_t^{eff} translate into labor productivity and measured TFP.

Marginal Product of Capital

The marginal product of capital tells us how much output rises when we add a tiny bit more capital, holding other inputs fixed. In other words, I compute MPK_t to isolate the

mechanical contribution of an input change given the technology level (see Solow 1957; Acemoglu 2009).

To calculate partial derivative of Y_t with respect to K_t, holding L_t constant, and (for the derivative) treating A_t^{eff} as constant at time t, I begin from $Eq.(2)$.

Treating A_t^{eff} and ${L_t}^{1-\alpha}$ as constants with respect to K_t:

$$\frac{\partial Y_t}{\partial K_t} = A_t^{eff} L_t^{1-\alpha} \cdot \frac{\partial}{\partial K_t}(K_t^{\alpha})$$

Applying the power rule:

$$\frac{\partial}{\partial K_t}(K_t^{\alpha}) = \alpha K_t^{\alpha-1}$$

I then substitute back in:

$$\frac{\partial Y_t}{\partial K_t} = A_t^{eff} L_t^{1-\alpha} \cdot \alpha K_t^{\alpha-1}$$

And reorder terms:

$$\frac{\partial Y_t}{\partial K_t} = \alpha A_t^{eff}\ K_t^{\alpha-1}\ L_t^{1-\alpha}$$

This yields a new equation for the marginal product of capital:

$$MPK_{\mathrm{t}} \equiv \frac{\partial Y_{\mathrm{t}}}{\partial K}{}_{\mathrm{t}} = \alpha A_t^{eff}\ {K_{\mathrm{t}}}^{\alpha-1}\ {L_{\mathrm{t}}}^{1-\alpha} \quad (3)$$

I can draw three conclusions from this exercise:

1. **Diminishing returns to capital are built in (given labor).** This is because $0 < \alpha < 1$, and $\alpha - 1 < 0$. So as K_t rises, $K_t^{\alpha-1}$ falls. In other words, holding L_t fixed, each additional unit of capital adds **less** extra output than the previous one, the standard diminishing marginal product logic (see Acemoglu 2009).

2. **AI represented through A_t^{eff} scales the MPK one-for-one.** A_t^{eff} is multiplicative in Eq. (3). If A_t^{eff} rises by (say) 10%, MPK rises by 10%, all else equal.
3. **More labor raises the productivity of capital.** Because $L_t^{1-\alpha}$ appears positively, MPK is higher when labor is larger. In other words, capital is more productive when there is more labor to work with it: machines need operators, and systems need people, etc.

The upshot is that if **AI increases** A_t^{eff}, then the **MPK rises**, which helps explain why firms may want to invest more, and why returns can be high for early adopters. However, I should also add that in this modelling exercise so far, markets are competitive. That means there are no markups, so that capital earns its marginal product. By extension, the rental rate (or real return to capital services) tracks MPK (see Acemoglu 2009).

Marginal Product of Labor

The marginal product of labor tells us how much output rises when we add a tiny bit more labor, holding other inputs fixed. In other words, I compute MPL_t to isolate the *mechanical contribution* of an input change given the technology level (see Solow 1957; Acemoglu 2009).

To calculate partial derivative of Y_t with respect to L_t, holding K_t constant, and (for the derivative) treating A_t^{eff} as constant at time t, I begin from $Eq.(2)$.

Treating A_t^{eff} and K_t^{α} as constants with respect to L_t. Then:

$$\frac{\partial Y_t}{\partial L_t} = A_t^{eff}\ K_t^{\alpha} \cdot \frac{\partial}{\partial L_t}(L_t^{1-\alpha})$$

Applying the power rule:

$$\frac{\partial}{\partial L_t}(L_t^{1-\alpha}) = (1-\alpha)L_t^{-\alpha}$$

I then substitute back in:

$$\frac{\partial Y_t}{\partial L_t} = A_t^{eff}\ K_t^{\alpha} \cdot (1-\alpha) L_t^{-\alpha}$$

And reorder terms:

$$\frac{\partial Y_t}{\partial L_t} = (1-\alpha) A_t^{eff} K_t^{\alpha}\ L_t^{-\alpha}$$

This yields a new equation for the marginal product of labor:

$$MPL_t \equiv \frac{\partial Y_t}{\partial L_t} = (1-\alpha)\, A_t^{eff}\ K_t^{\alpha}\ L_t^{-\alpha} \quad (4)$$

And because $Y_t = A_t^{eff}\ K_t^{\alpha}\ L_t^{1-\alpha}$, we can rewrite $Eq.(4)$ as:

$$MPL_t = (1-\alpha)\frac{Y_t}{L_t} \quad (5)$$

Allowing me to thus connect the model to wages and labor productivity in plain language.

I can draw three conclusions from this exercise:

1. **Diminishing returns to labor are built in (given capital).** This is because $0 < \alpha < 1$; the exponent on labor in $Eq.(4)$ is $-\alpha < 0$. So as $L_t$rises, $L_t^{-\alpha}$falls. In other words, holding $K_t$fixed, the additional output produced by one more unit of labor **declines** as labor becomes more abundant.
2. **Capital deepening raises labor's marginal product.** $Eq.(4)$ rises with K_t^{α}. More capital raises MPL. Therefore, workers are more productive when they have more (or better) capital to work with—machines, tools, software, etc.
3. **AI technology (through A_t^{eff}) raises labor productivity mechanically.** Just as with MPK, the entire expression for MPL is multiplied by A_t^{eff}. So, if AI increases A_t^{eff}, then **MPL rises one-for-one**, all else equal. In other words, AI increases the productivity of labor by scaling output for given inputs.

The upshot is that if **AI increases** A_t^{eff}, then **the MPL rises**, which shifts labor demand outward. All else equal, firms *can* profitably employ more labor at a given wage, especially when AI complements workers and expands output (Acemoglu 2009)—albeit, equilibrium employment rises only if wages don't rise enough to offset the shift, and if the firm can sell more output profitably. That brings me to the fact that there is **a connection between this formulation and wages.** Because in this treatment I am assuming competitive labor markets, **real wages track MPL**: if labor is paid its marginal product, then $w_t = MPL_t$.

I can now move onto exploring **how output changes with respect to technology** and how that links to labor productivity and measured TFP.

Technology's Role in Bolstering Productivity

Recall that in this macroeconomic model technology is a multiplicative factor that scales how productively inputs are used (see Solow 1957; Acemoglu 2009; Syverson 2011). Given the Cobb Douglas production function, the **"marginal product of technology,"** MPA_t, is the additional output produced by a small increase in effective technology, **holding** K_t **and** L_t **fixed**. Formally:

$$MPA_t \equiv \frac{\partial Y_t}{\partial A_t^{eff}}$$

This formulation tells us how strongly output responds—mechanically—to improvements in A_t^{eff}. However, this is not a "factor payment" in the same way as wages or rental rates—firms don't literally pay "technology" its marginal product. To calculate this, I can again take derivatives by starting with $Eq.(2)$:

$$Y_t = A_t^{eff}\, K_t^{\alpha}\, L_t^{1-\alpha}$$

Holding K_t and L_t constant, I treat the term $K_t^{\alpha} L_t^{1-\alpha}$ as a constant with respect to A_t^{eff}. Then:

$\frac{\partial Y_t}{\partial A_t^{eff}} = \frac{\partial}{\partial A_t^{eff}} \left(A_t^{eff}\ K_t^{\alpha}\ L_t^{1-\alpha} \right) = K_t^{\alpha} L_t^{1-\alpha}$ because the derivative of A_t^{eff} with respect to itself is 1.

Therefore:

$$MPA_t \equiv \frac{\partial Y_t}{\partial A_t^{eff}} = K_t^{\alpha} L_t^{1-\alpha} \quad (6)$$

I can draw two conclusions from this exercise:

1. The payoff to raising A_t^{eff} is larger when the economy (or sector) has a larger effective bundle of inputs.
2. Because A_t^{eff} multiplies the whole production function, it also multiplies both marginal products: MPK_t rises proportionally with A_t^{eff} $(Eq. 3)$; MPL_t rises proportionally with A_t^{eff} $(Eq. 4)$.

Finally, using $Eq. (2)$, since $Y_t = A_t^{eff}\ K_t^{\alpha}\ L_t^{1-\alpha}$, we can rewrite $Eq. (6)$ as:

$$MPA_t = \frac{Y_t}{A_t^{eff}} \quad (7)$$

This makes the economic meaning intuitive: **a 1% increase in A_t^{eff} implies a 1% increase in Y_t** (holding K_t, L_t fixed), because technology enters multiplicatively. And now that I've pinned down how output responds to A_t^{eff}, the next natural step is to connect the framework to: **Labor productivity** (Y_t/L_t) and **Measured TFP** $(TFP_t \equiv Y_t/(K_t^{\alpha} L_t^{1-\alpha}))$, which will set us up to complicate the model with technological standardization and diffusion dynamics.

Labor Productivity

First, I connect the model to **what productivity statistics are usually reporting**: labor productivity measured as output per unit of labor input, usually worker output per hour worked. Formally, starting with $Eq.(2), Y_t = A_t^{eff} K_t^{\alpha} L_t^{1-\alpha}$, I can divide both sides by L_t:

$$\frac{Y_t}{L_t} = A_t^{eff} K_t^{\alpha} \frac{L_t^{1-\alpha}}{L_t}$$

If I simplify the labor term:

$$\frac{L_t^{1-\alpha}}{L_t} = L_t^{1-\alpha-1} = L_t^{-\alpha}$$

That translates to:

$$\frac{Y_t}{L_t} = A_t^{eff} K_t^{\alpha} L_t^{-\alpha}$$

Finally, I can collect terms as a capital–labor ratio:

$$K_t^{\alpha} L_t^{-\alpha} = \left(\frac{K_t}{L_t}\right)^{\alpha}$$

And this yields:

$$\frac{Y_t}{L_t} = A_t^{eff} \left(\frac{K_t}{L_t}\right)^{\alpha} \quad (8)$$

I can draw two conclusions from this exercise.

1. Labor productivity rises for two distinct reasons: **technology/effectiveness:** A_t^{eff} and **capital deepening:** $(K_t/L_t)^{\alpha}$. Therefore, **the clean "AI channel" in this benchmark is** $\boldsymbol{A_t^{eff}}$. Mechanically, if AI improves A_t^{eff}, then Y_t/L_t rises **one-for-one** with A_t^{eff}, holding K_t/L_t fixed.

2. **Labor productivity improvements tie back to wages.** From the earlier identity ($Eq.5$) I already established: $MPL_t = (1-\alpha)\frac{Y_t}{L_t}$. So, when labor productivity rises, the marginal product of labor rises proportionally.

Finally, a useful **"growth rate" interpretation** follows: If I take log changes over time, $Eq.(8)$ implies labor productivity growth: ≈ growth in $A^{eff} + \alpha \times$ growth in capital per worker.

I can now turn to measured Total Factor Productivity and show that in this setup) **TFP is exactly the effective technology term A_t^{eff}**.

Total Factor Productivity

If I again start from the benchmark production function ($Eq.2$), $Y_t = A_t^{eff} K_t^{\alpha} L_t^{1-\alpha}$, and divide both sides by $K_t^{\alpha} L_t^{1-\alpha}$:

$$\frac{Y_t}{K_t^{\alpha} L_t^{1-\alpha}} = \frac{A_t^{eff} K_t^{\alpha} L_t^{1-\alpha}}{K_t^{\alpha} L_t^{1-\alpha}}$$

I can then cancel like terms:

$$\frac{Y_t}{K_t^{\alpha} L_t^{1-\alpha}} = A_t^{eff}$$

Therefore, **measured TFP equals the effective technology term**. In other words:

$$TFP_t \equiv \frac{Y_t}{K_t^{\alpha} L_t^{1-\alpha}} = A_t^{eff} \quad (9)$$

I can draw two conclusions from this exercise:

1. **AI shows up in TFP by construction**. Since A_t^{eff} multiplies the whole production function, it acts as a "Hicks-neutral" shifter (see Acemoglu 2009), meaning it increases output (and the marginal product of all inputs) without requiring any increase in the quantity of capital or labor employed. **So, if AI raises A_t^{eff}** (i.e., it truly makes the economy more effective at combining inputs), **measured TFP rises.**

2. **But, because TFP is a residual, it can include more than "pure technology".**

   Therefore, TFP can reflect: technology and organizational change, measurement error in outputs and inputs (see Gordon 2016; Brynjolfsson et al. 2019); mismeasured intangible capital such as software, data, and organizational capital (see Brynjolfsson et al. 2021); variations in utilization (e.g., capacity utilization), and compositional shifts that aren't fully captured in $K$and L (Acemoglu 2009).

The upshot is if AI-related intangible investments are not fully captured in measured K_t, then some of the payoff may appear in TFP rather than in "capital deepening." Conversely, if output quality improvements are hard to measure (better services, fewer errors, faster turnaround), measured Y_t may understate true output—making TFP look smaller than the lived experience inside firms.

Incorporating Frictions: The Standardization and Assimilation Wedge

The baseline model above assumes that any increase in frontier AI capability captured in A_t^{eff} is instantly and costlessly available to the entire economy. However, the history of General Purpose Technologies—from electricity to the internet—demonstrates that this is rarely the case. As Chapter 2 and 10 argue, the gap between invention and impact is defined by the hard work of standardization and organizational adaptation: The frontier can advance quickly (spectacular demos, rapid model improvements), while the effective technology embedded in real production processes advances slowly. To capture this reality, I now relax the assumption of perfect AI diffusion across the entire economy.

Defining the Enhanced Model

While retaining the underlying framework outlined in the baseline model outlined above, I can complicate the effective technology term, A_t^{eff}. Let $A_t^{frontier}$ denote the **best available**

technology/AI capability at time t— what is technically achievable by leading firms at the frontier, given the best models, best data infrastructure, best workflows, best talent, and so on. I can **connect the frontier to the "effective" technology used in production** by introducing a wedge λ_t that captures the idea that **not all firms, sectors, and workflows can (yet) use the frontier effectively** — because of adoption lags, a lack of complementary investments, standardization gaps and similar frictions, a training deficit, and the costs of organizational rewiring and integration. Formally:

$$A_t^{eff} = \lambda_t A_t^{frontier}, 0 < \lambda_t \leq 1 \quad (10)$$

This equation implies two things:

1. **Breadth:** λ_t can be read as "the fraction of production that is effectively operating near the frontier," or "the share of firms/sectors/workflows that have successfully embedded AI into production." If λ_t = 1, it means 100% of AI frontier capability is being used. (perfect efficiency). Conversely, if λ_t = 0.1, it means only 10% of the capability is getting through (high friction).
2. **Effective intensity:** more generally, λ_t is a *summary statistic* for "how close the economy is to turning frontier capability into routine practice."

That upshot is that λ_t is an open-ended term that flexibly absorbs many real frictions. These include the lack of standards or "proto-standards" that prevent AI systems from talking to one another; fragmented data architectures that trap information in legacy silos rather than feeding it into models ("missing plumbing"); the absence of industrial-grade pipelines to monitor and maintain models in production ("MLOps"); and the sheer difficulty of retooling workflows, management practices, and worker skills to accommodate non-human agents.

TFP and Labor Productivity.

TFP in this model is defined as "frontier AI technology multiplied by its diffusion." Formally, I can start with $Eq.(9)$, $TFP_t = A_t^{eff}$, and substitute $Eq.(10)$ into $Eq.(9)$:

$$TFP_t = A_t^{eff} = \lambda_t A_t^{frontier} \quad (11)$$

I can draw an important conclusion from this equation:

- Even if **frontier AI improves rapidly** (rising $A_t^{frontier}$), *TFP can remain modest if diffusion/standardization remains incomplete* ($\lambda_t < 1$).

Labor productivity is also governed by "frontier AI technology multiplied by its diffusion." Formally, I can start from $Eq.(8)$, which I already derived:

$$\frac{Y_t}{L_t} = A_t^{eff} \left(\frac{K_t}{L_t}\right)^{\alpha}$$

If I substitute $Eq.(10)$, $A_t^{eff} = \lambda_t A_t^{frontier}$:

$$\frac{Y_t}{L_t} = \lambda_t A_t^{frontier} \left(\frac{K_t}{L_t}\right)^{\alpha} \quad (12)$$

I can draw an important conclusion from this equation:

- Labor productivity growth is jointly driven by three things. They are frontier AI capability ($A_t^{frontier}$); diffusion/standardization (λ_t); and capital deepening (K_t/L_t) —the accumulation of more capital per worker.

The upshot is that **for AI to fully improve economy-wide productivity** operationalized as TFP and labor productivity, economic actors must increase their effective adoption rate of the technology, λ_t. In practice, closing the gap between potential and realized technology requires successfully spreading "plug and play" standards, rewiring workflows, training workforces, and integrating more seamless data and models into organizations. Conversely, if that does not happen and λ_t stays low for years, measured TFP can lag even while frontier capability leaps

forward, and workers may not become more observably productive.

Modeling the AI Adoption Trajectory

The pieces are in place for me to now **specify an S-curve diffusion dynamic for λ_t**, a logistic progression typical of major industrial shifts (Lipsey et al. 2005). Specifically, I can deduce a diffusion law for λ_t from $eq.(10)$ that where $A_t^{frontier}$ grows slowly at first, then rises rapidly, then plateaus. In the initial **convex phase**, adoption is an expensive form of R&D restricted to "frontier firms"—those with the capital buffers to tolerate high failure rates and the technical talent to integrate bespoke systems. The inflection point arrives only when standardization lowers the unit cost of integration, transforming the technology from a risky experiment into a competitive necessity. As "best practices" calcify and knowledge spillovers occur, the curve accelerates vertically. Eventually, however, the process enters a **concave saturation phase**: the pool of potential adopters shrinks to a "long tail" of laggards—entities constrained by rigid regulatory frameworks, capital immobility, or incompatible legacy infrastructure that makes retrofitting prohibitively costly.[268]

[268] I hasten to emphasize that, while the narrative above describes endogenous economic choices (e.g., firms rationally waiting for standardization to lower costs), this does not require formally endogenizing λ_t via an optimization problem (where agents choose specific investment levels to maximize intertemporal profit). Instead, following the tradition of the technology diffusion literature (e.g., Griliches 1957; Mansfield 1961; Comin and Hobijn 2010), I treat the logistic S-curve as a reduced-form structural representation of these complex underlying dynamics. This approach allows the model to capture the *aggregate trajectory* of AI adoption—and the specific

Formally, the growth rate for λ_t is proportional to both the current adoption level λ_t, and the remaining adoption potential $(1-\lambda_t)$, and I can model this in **continuous time** to cleanly map this dynamic onto a logistic curve:[269]

$$\frac{d\lambda_t}{dt} = \kappa\,\lambda_t(1-\lambda_t), \kappa > 0 \quad (13)$$

Where κ is the **speed of diffusion** and a higher κ means standards, tools, training pipelines, and "plug-and-play" deployment practices spread quickly. Conversely, a lower κ means diffusion is bottlenecked by integration costs, governance, skills, compliance, legacy IT, and organizational inertia.

I can now solve Eq. (13) to show it generates an S-curve. Starting from Eq. (13) stated in continuous time:

$$\frac{d\lambda}{dt} = \kappa\lambda(1-\lambda)$$

"standardization wedge" frictions—without the intractability of modeling heterogeneous firm-level decisions across every sector of the economy.

[269] While I model this in continuous time for analytical tractability, it is worth noting that the literature often utilizes discrete time specifications for two distinct reasons. First, empirical studies (e.g., Geroski 2000) favor discrete time because economic data is reported in fixed intervals (quarters or years). Second, many theoretical treatments (e.g., Brock and Durlauf 2001) utilize discrete time to explicitly model the sequential nature of decision-making, where agents observe past outcomes before committing to future adoption. However, the continuous approach used here offers cleaner closed-form solutions for the transition dynamics and allows for more intuitive comparative statics regarding the speed of convergence, without altering the fundamental logistic shape of the trajectory.

I can separate the variables:

$$\frac{d\lambda}{\lambda(1-\lambda)} = \kappa\, dt$$

And then use partial fractions by exploiting a standard identity:

$$\frac{1}{\lambda(1-\lambda)} = \frac{1}{\lambda} + \frac{1}{1-\lambda}$$

So:

$$\left(\frac{1}{\lambda} + \frac{1}{1-\lambda}\right) d\lambda = \kappa\, dt$$

I can then integrate both sides:

$$\int \left(\frac{1}{\lambda} + \frac{1}{1-\lambda}\right) d\lambda = \int \kappa\, dt$$

$$\ln(\lambda) - \ln(1-\lambda) = \kappa t + C$$

where C is a constant of integration.

And solve for $\boldsymbol{\lambda(t)}$ by first combining logs:

$$\ln\left(\frac{\lambda}{1-\lambda}\right) = \kappa t + C$$

Followed by exponentiating:

$$\frac{\lambda}{1-\lambda} = e^{\kappa t + C}$$

Let $e^C \equiv B > 0$; I can then write:

$$\frac{\lambda}{1-\lambda} = Be^{\kappa t}$$

And solve for λ:

$$\lambda(t) = \frac{Be^{\kappa t}}{1 + Be^{\kappa t}} = \frac{1}{1 + B^{-1}e^{-\kappa t}}$$

Finally, it is conventional to re-parameterize B in terms of the **inflection point** t_0, the time at which $\lambda = 1/2$, i.e., $\lambda_{t_0} = 0.5$, which yields the following equation:[270]

$$\lambda_t = \frac{1}{1+e^{-\kappa(t-t_0)}} \quad (14)$$

I can draw three conclusions from this equation:

1. For $t \ll t_0$, $\lambda_t$is small (early adoption by a minority of "frontier firms").
2. Around $t \approx t_0$, diffusion is fastest (the steep part of the S-curve).
3. For $t \gg t_0$, λ_t approaches 1 (late-stage saturation).

The upshot is that **productivity can look flat and then suddenly surge.**

Moreover, recall that, as stated by Eq. 11, TFP under S curve diffusion is $TFP_t = \lambda_t A_t^{frontier}$. Substituting the logistic diffusion path (Eq. 14) therefore gives:

$$TFP_t = \lambda_t A_t^{frontier} = \frac{A_t^{frontier}}{1+e^{-\kappa(t-t_0)}} \quad (15)$$

I can draw an important conclusion from this equation:

- Even if $A_t^{frontier}$ is rising (improving models), measured TFP can remain modest for a while because, early on, $\lambda_t$is small (adoption is limited to leading firms), so the economy only "turns on" a small share of frontier capability. Then, as standards and complements

[270] It is important to note that introducing t_0 does not imply a shift to discrete time steps. Instead, t_0 is simply a scalar constant representing the specific continuous time coordinate (e.g., $t_0 = 2028.4$) where the adoption curve intersects the midpoint λ = 0.5). The variable t remains continuous, and $Eq.(14)$ defines a smooth, differentiable trajectory for every instant of the transition, not just at fixed intervals. Re-parameterizing in this way merely makes the equation more interpretable by centering the diffusion process around its period of maximum acceleration.

arrive (raising κ effectively, or moving the curve forward), λ_t accelerates and **TFP growth can appear to "arrive late."** (see Crafts 2021; Brynjolfsson et al. 2021).

There is a larger upshot: in this book I have documented a perplexing puzzle. While I've chronicled the exponential increases in investment in AI, exponential improvements in performance, exponential reductions in its cost, and exponential adoption patterns, I have also summoned considerable evidence about AI's disappointing results and a slowing down in its uptake and impact. From the perspective of the logistic curve that expresses how GPTs evolve, this is not surprising. To see why, let's start from $Eq.(13)$, $\frac{d\lambda_t}{dt} = \kappa\lambda_t(1-\lambda_t)$: If $\lambda_t$is small, then $(1-\lambda_t) \approx 1$. So, $Eq.(13)$ essentially becomes: $\frac{d\lambda_t}{dt} \approx \kappa\lambda_t$ and this implies approximately:

$$\lambda_t \approx \lambda_{t_i} e^{\kappa(t-t_i)} \tag{16}$$

So that adoption must eventually slow as it approaches saturation (as the λ_t term shrinks toward zero): **the last firms are hardest to convert**, and there are fewer non-adopters left.[271]

Revisiting Productivity

[271] We can see this mathematically by analyzing the percentage growth rate of technological adoption. From Eq. (13), we can derive $\frac{d\ln\,\lambda_t}{dt} = \kappa(1-\lambda_t)$. This identity reveals the precise mechanics of the slowdown: when adoption (λ_t) is near zero, the percentage growth rate is roughly constant at κ (the "hype" phase). However, as adoption spreads, the growth rate does not stay constant; it decays linearly with the saturation level. As $\lambda_t \to 1$, **the growth rate of adoption approaches zero**, therefore ushering in the end of the exponential regime.

We are now ready to tie both labor productivity and TFP into these logistic accented dynamics.

First, I can now **plug the S-curve directly into labor productivity**, substituting $Eq.(14)$ into $Eq.(12)$:

$$\frac{Y_t}{L_t} = \frac{A_t^{frontier}}{1+e^{-\kappa(t-t_0)}}\left(\frac{K_t}{L_t}\right)^{\alpha} \quad (17)$$

I can draw two conclusions from this equation:

1. Even if frontier AI capability $A_t^{frontier}$ is high (spectacular demos), the economy only realizes a fraction of it until diffusion and standardization progress.
2. The diffusion term λ_t creates the classic **slow-then-fast-then-slow** pattern typical of GPT adoption: early productivity gains are concentrated in leading firms, then spread broadly only after complements (standards, workflows, training, governance) reduce the "transaction costs" of adoption.

Second, I can derive the growth rate of labor productivity under these diffusion dynamics. From $Eq.(12)$, $\frac{Y_t}{L_t} = \lambda_t A_t^{frontier}\left(\frac{K_t}{L_t}\right)^{\alpha}$, I can take logs and differentiate:

$$\frac{d\ln(Y_t/L_t)}{dt} = \frac{d\ln\lambda_t}{dt} + \frac{d\ln A_t^{frontier}}{dt} + \alpha\frac{d\ln(K_t/L_t)}{dt}$$

Now I can substitute $\frac{d\ln\lambda_t}{dt} = \kappa(1-\lambda_t)$ to arrive at:

$$\frac{d\ln(Y_t/L_t)}{dt} = \kappa(1-\lambda_t) + \frac{d\ln A_t^{frontier}}{dt} + \alpha\frac{d\ln(K_t/L_t)}{dt} \quad (18)$$

I arrive at two related conclusions from this exercise:

1. **During the diffusion phase**, labor productivity growth can be high because $\kappa(1-\lambda_t)$ is sizeable.
2. **Later**, once diffusion saturates ($\lambda_t \to 1$), that term fades, and labor productivity depends more on frontier improvements, and capital deepening.

The upshot is powerful: Even if the **Capital Deepening** term, $\alpha \frac{d \ln (K_t/L_t)}{dt}$, is huge (because of huge spending on AI chips and data centers documented in Chapter 10) and the **Frontier** term is huge (because of the amazing progress made on AI training and inference), if the **Standardization Wedge** ($\kappa(1-\lambda_t)$ is low, the sum remains low: the growth rate of labor productivity may remain relatively stagnant.

Third, I can derive the **growth rate of measured TFP under these diffusion dynamics.**

Recall $Eq.(11)$: $TFP_t = \lambda_t A_t^{frontier}$. If I take logs and differentiate:

$$\frac{d \ln\ TFP_t}{dt} = \frac{d \ln\ \lambda_t}{dt} + \frac{d \ln A_t^{frontier}}{dt}$$

Now I can substitute $\frac{d \ln\ \lambda_t}{dt} = \kappa(1-\lambda_t)$ to arrive at:

$$\frac{d \ln\ TFP_t}{dt} = \kappa(1-\lambda_t) + \frac{d \ln\ A_t^{frontier}}{dt} \quad (19)$$

I can conclude two important insights into the contributors to TFP growth from this equation:

1. **Frontier advances are insufficient on their own:** Because Frontier Growth is driven by how fast the best available capability improves and Diffusion Growth is driven by how quickly that capability becomes routine across the economy, big frontier advances don't automatically translate into TFP until λ_t rises to an appreciable degree.
2. **The "Pent-Up" TFP Boom:** The term $\kappa(1-\lambda_t)$ implies that TFP growth from diffusion is not constant—it peaks when adoption is accelerating fastest (the steep part of the S-curve). This resolves the current paradox I outlined towards the start of this chapter: in 2025 we were in the flat, early part of the curve where λ_t is small. The model predicts that the surge in measured TFP is not missing, but backloaded; it will arrive only when

standardization (κ) unlocks widespread adoption; if we look back to past industrial revolutions, however, that process may take decades to reach fruition.

EXTENDING DIFFUSION FRAMEWORK TO PROFITS, WAGES & DISTRIBUTION

Up to this point, this modeling exercise has treated the economy as essentially competitive: factors are paid their marginal products ($w_t = MPL_t$), and, with constant returns to scale, ($(\alpha + (1 - \alpha) = 1)$ output is "exhausted" by payments to capital and labor, the sum of those payments equals exactly total output ($MPL \cdot L + MPK \cdot K = Y$). Therefore, pure economic profits are zero.

However, it does not explain the *distribution* of the value that's created. The model struggles hitherto to say anything incisive about **profits** and **who captures AI's gains**. The good news is that to get traction on profits and distributional dynamics—and without building a full endogenous growth model—we only need one additional ingredient: a **markup/rent wedge** that captures market power and temporary (or persistent Ricardian) rents from AI advantages.

Defining the Markup

While I have already established and embedded the **standardization friction** $\boldsymbol{\lambda_t}$, which limits how much *output* is produced, I now introduce a second friction, the "**Markup Wedge" ($\boldsymbol{\mu_t}$)**, which limits how that output is *distributed* and priced. To capture this friction, I adopt the standard framework from the industrial organization literature that models market power as a wedge between price and marginal cost (Hall 1988). This approach has been used by De Loecker

et al. (2020) and Barkai (2020) to explain the structural decline in the labor share (see also Rotemberg and Woodford 1992).[272] Formally:

$$\mu_t \equiv \frac{P_t}{MC_t}, \mu_t \geq 1 \quad (20)$$

Where:

- P_t is the output price.
- $MC_t$is marginal cost.
- If $\mu_t = 1$, we're back in the competitive benchmark.
- If $\mu_t > 1$, firms have **market power / rents** and some of the gains from productivity need not be passed through to consumers as lower prices or to workers as higher wages.

While this wedge allows me to introduce profits, "winner-takes-all", and "winner-take-most" outcomes, it behooves me to clarify that μ_t does not imply that AI companies are necessarily behaving like traditional monopolists—restricting output to gouge consumers. Indeed, in the digital economy, this wedge frequently represents **Ricardian Rents** (see Bessen and Calligaris 2018; Galetovic 2021; Gochberg and Menaldo 2022). As I argued in Chapter 4, these types of rents arise when a firm possesses a scarce, highly productive asset—

[272] Hall (1988) was the first to show that under perfect competition, productivity growth should track price declines perfectly; persistent residuals therefore imply a markup wedge. This framework was revitalized by empirical work decades later: De Loecker et al. (2020) use it to document that aggregate markups in the US rose from approximately 1.2 in 1980 to 1.6 during the 2010s. Similarly, Barkai (2020) employs this logic to demonstrate that a decline in the labor share during the same period is mathematically driven by a rising "profit share" that crowds out returns to both labor and capital.

such as a proprietary foundation model, a unique dataset, or an entrenched ecosystem—that allows it to produce at a significantly lower marginal cost (or higher quality-adjusted value) than its competitors.

To see this formally, consider a "Superstar" AI firm. In a competitive market, the market price (P_t) is often set by the "marginal firm"—a less efficient competitor using generic tools (see Bain 1956; Milgrom and Roberts 1982). The Superstar firm, leveraging its unique data moat or superior IP, operates at a much lower marginal cost ($MC_{superstar}$). Specifically:

- For the marginal firm, $P_t \approx MC_marginal$, so $\mu \approx 1$.
- For the Superstar firm, $P_t \gg MC_superstar$, so $\mu \gg 1$.

In this scenario, as I outlined in Chapter 4, the firm is not artificially inflating prices; it is simply capturing the value of its scarcity.

However, for our purposes here, the macroeconomic effect is identical to market power: the "Profit Share" of income rises, while the share of revenue going to labor and standard capital falls. Therefore, when we model $\mu_t > 1$, we are capturing the **returns to scarcity** in the AI supply chain. Whether termed "market power" or "efficiency rents," the outcome is a concentration of economic gains in the hands of those who own the bottleneck assets.

How markups change factor payments

Under imperfect competition, firms choose inputs where **marginal revenue product** equals input cost. A convenient and standard reduced-form implication is: factor prices equal marginal products **discounted by the markup**. This is widely used in empirical industrial organization and macro discussions of markups (Hall 1988; De Loecker and Eeckhout 2017).

Formally, we can write the equilibrium conditions for wages (w_t) and the rental rate of capital (r_t) as:

$$w_t = \frac{1}{\mu_t} MPL_t, r_t = \frac{1}{\mu_t} MPK_t \quad (21)$$

I can conclude from this equation that:

- Even if AI raises **true marginal products** (MPL and MPK) through A_t^{eff}, **observed wages and rental rates can rise less** if μ_t rises, because a larger fraction of the gain is captured as rents/profits.

The upshot is that during the AI era a **Decoupling Mechanism** may arise even if AI drives the Marginal Product of Labor (MPL_t) sky-high through rapid technological progress ($A_t^{frontier}$): observed real wages (w_t) will fail to keep pace if the Markup Wedge (μ_t) is rising simultaneously. In other words, $Eq.(21)$ shows that μ_t acts as a divisor—effectively shrinking the share of productivity that flows to workers. This mathematically formalizes a "hollowed out" growth thesis that is often repeated in the media and by pundits and populist politicians: while the pie is getting bigger ($higher\ MPL$), the slice going to labor is getting thinner ($higher\ \mu$).

However, this equation also reveals a powerful alternative path—a "Race to the Top" scenario.

First, consider **Absolute Real Wages**. Even if the labor *share* is compressed by a rising markup (μ_t), the *level* of real wages (w_t) can still skyrocket if the underlying productivity term (MPL_t) expands fast enough. If the standardization friction is overcome quickly ($rapid\ rise\ in\ \lambda_t$), the resulting boom in effective technology ($A^{eff}{}_t$) can overpower the markup drag. In this scenario, workers may receive a smaller slice of the pie, but the pie itself grows so explosively that they are still better off in absolute terms.

Second, consider the **Labor Share** itself. The markup wedge is not a constant of nature; it is a variable determined by market structure. The "Decoupling" scenario assumes that μ_t rises indefinitely. But if the diffusion of AI leads to "democratization"—where barriers to entry fall

and competition intensifies—the standardization process ($\lambda_t \to 1$) can act as a force that eventually narrows the markup wedge. In this optimistic case, as AI becomes ubiquitous and "plug-and-play," temporary, “novelty” rents are competed away before they metastasize into the Ricardian version (more on this below), μ_t falls back toward 1, and the share of income flowing to labor recovers.

The ultimate impact on the workforce, therefore, depends on the speed of this race: does the **Diffusion Wedge** (which grows the pie) narrow faster than the **Markup Wedge** (which shrinks the slice) expands?

Visualizing How This Wedge Separates the Worker from their Output

I can substitute the Cobb-Douglas definition of MPL $\left(MPL_t = (1-\alpha)Y_t\frac{}{L}_t\right)$ directly into Equation (21). This yields a crystal-clear **"Payoff Line"** for the AI economy:

$$w_t = \left[\frac{(1-\alpha)}{\mu_t}\right] \cdot \left(\frac{Y_t}{L_t}\right) \tag{22}$$

This identity dictates that real wages are simply Labor Productivity $\left(\frac{Y_t}{L_t}\right)$ scaled by the labor share parameter $(1-\alpha)$ and deflated by the markup (μ_t).

This simple equation reveals the precise mechanics of the trap:

- If the markup μ_t remains constant (as in perfect competition), wages and rental rates grow at the exact same percentage rate as productivity. They differ in level, but they share the same trend.
- If the AI era causes markups to rise ($\mu_t \uparrow$) due to "Winner-Take-Most" dynamics, both wages and rental rates will mathematically rise more slowly than productivity.

- The gap between the "Productivity Curve" (what workers produce) and the "Wage Curve" (what workers earn) widens—and that missing value flows directly into the Profit Share: even if AI drives Labor Productivity $\left(\frac{Y_t}{L_t}\right)$ sky-high via the "Scale Effect," observed real wages (w_t) will fail to keep pace if the "Gatekeepers" simultaneously expand the Markup Wedge.

This derivation allows us to move beyond vague political rhetoric and pinpoint the precise economic mechanism behind the popular feeling that the system is "rigged."[273] Indeed, critics of the creative destruction process often describe technological change and the digital economy as emblematic of "Vampire Capitalism"—a system where financial and technological elites extract value from the real economy (Kennedy 2017), rather than creating it. $Eq.(22)$ reveals that this is not just a metaphor; it is a measurable structural condition: When μ_t rises, the firm acts as a "rentier" or vampire. It inserts a wedge into the production process that severs the link between what a worker *does* and what they *get* (their stagnant wage, w_t). The difference does not disappear; it is transferred directly to the owners of the scarcity.

This leads to three stark conclusions about the distribution of AI wealth:

[273] This mechanism bridges two strands of economic literature. The “decoupling” of wages via markups builds on the work of De Loecker et al. (2020) and Barkai (2020), who document how rising market power has depressed the labor share globally. However, I extend this logic to the specific context of AI automation (following Acemoglu and Restrepo 2018), demonstrating that high markups do not just lower the labor share—they short-circuit the 'Scale Effect' required to maintain employment levels.

1. If AI delivers a "productivity miracle" by doubling $\left(\frac{Y_t}{L_t}\right)$, but gatekeepers simultaneously double their markups to capture the value (μ_t doubles), the real wage (w_t) will remain mathematically flat. The workers will have produced twice as much wealth yet captured *none* of it. The entire efficiency gain is effectively "plundered" by the markup wedge.
2. In a healthy competitive market, profits are a **reward** for efficiency. In a "Gatekeeper" market (high μ_t), profits are a **tax** on the rest of the economy. This distinction is crucial: The high margins of AI hyperscalers are not merely a signal of their genius; they are a measure of the "tax" they are levying on every downstream application, developer, and worker in the ecosystem.
3. Finally, this model proves that inequality is not an inevitable byproduct of technology. The variable μ_t is not a physical constant like gravity and the "decoupling" of wages is not a technical necessity of AI; it is a variable determined by **market power**.

Figuring Out Profits

To score this race effectively, therefore, we must look at the Profit Share. I can derive **profits directly (and transparently)** by defining them as revenue minus factor payments (in real terms, you can think of $P_t = 1$ for exposition or treat Y_t as real value added). This feeds into the following equation:

$$\Pi_t \equiv Y_t - w_t L_t - r_t K_t \tag{23}$$

I can now substitute the markup-adjusted factor prices (mechanics):

From $Eq.(21)$:

$$w_t L_t + r_t K_t = \frac{1}{\mu_t}(MPL_t L_t + MPK_t K_t)$$

Under constant returns to scale (true in Cobb–Douglas), Euler's theorem on homogenous functions implies:

$$MPL_t L_t + MPK_t K_t = Y_t$$

So:

$$w_t L_t + r_t K_t = \frac{1}{\mu_t} Y_t$$

Plugging into $Eq.(23)$, the profit share under markups:

$$\Pi_t = \left(1 - \frac{1}{\mu_t}\right) Y_t \tag{24}$$

I can draw three conclusions from this exercise:

- **Profits are zero in the competitive benchmark** ($\mu_t = 1$),
- **profits rise when market power/rents rise** ($\mu_t > 1$),
- AI can influence profits through **two channels**: raising output Y_t via productivity; changing μ_t via market concentration, IP, platforms, switching costs, and "winner-take-most" dynamics.

The upshot is that we have mathematically formalized the "Superstar Economy." This derivation confirms that rising industrial concentration or persistent Ricardian advantages do not simply shift money between capital and labor—the traditional concern of automation anxiety—but actually siphon value away from *both* factors into a distinct "profit bucket." This provides the theoretical scaffolding necessary to evaluate the empirical trends observed by Barkai (2020) and others: a world where GDP grows, but the combined share paid to workers and machine-owners shrinks because the markup wedge has effectively monetized the scarcity of the new technology.

I hasten to emphasize, however, that we must recognize that this derivation and the supporting empirical evidence represent a backward-looking analysis of the *Third* Industrial

Revolution. The rising markups and declining labor shares documented by researchers explain how the rise of the PC, the internet, and early digital platforms concentrated value over the last forty years. They do not, however, guarantee how the *Fourth* Industrial Revolution will unfold. To extrapolate these mechanics forward into the AI era, we cannot rely on historical inertia alone. Instead, the ultimate trajectory of the labor and profit shares will depend on the interplay of three evolving factors: **diffusion dynamics** (specifically, whether AI spreads fast enough to trigger the "Race to the Top"), the unpredictable nature of **creative destruction** (whether AI acts as a "great equalizer" that erodes incumbent moats or a centralizing force that fortifies them), and the decisive role of **regulation** (which has the power to either entrench superstar advantages or forcefully narrow the markup wedge).

Transitioning to Modeling the Distribution of Value Created

However, there are still some steps to go before I can make definitive statements about how market gains are doled out. I must first make distribution transparent by identifying the labor share, capital share, and profit share explicitly.

Using Cobb–Douglas properties:

- $MPL_t L_t = (1-\alpha)Y_t$
- $MPK_t K_t = \alpha Y_t$

and $Eq.(21)$, I can arrive at factor shares with markups:

$$\frac{w_t L_t}{Y_t} = \frac{1-\alpha}{\mu_t}, \frac{r_t K_t}{Y_t} = \frac{\alpha}{\mu_t}, \frac{\Pi_t}{Y_t} = 1 - \frac{1}{\mu_t} \quad (25)$$

I can draw two valuable conclusions from this exercise:[274]

[274] It is important to note the accounting implication of this result. In a competitive model ($\mu_t = 1$), the labor and capital shares sum to 1 ($\alpha + (1-\alpha) = 1$). However, in this imperfectly

1. If μ_t rises, the **labor share falls** and **profit share rises**, even if productivity is rising.
2. If μ_t falls (because competition increases, for example), more of the productivity gain flows to **lower prices and higher real incomes** across the economy.

The upshot is that several of the empirical patterns I discussed above, the declining labor share and rising profits for some periods/sectors (see Barkai 2020; Autor et al. 2020), while difficult to explain with a purely competitive Cobb–Douglas framework, flow naturally from a treatment that incorporates markups/rents as a wedge or friction.

Interestingly, this framework also reveals a counterintuitive dynamic regarding capital:

Notice that a rising markup (μ_t) suppresses the capital share $\left(\frac{\alpha}{\mu_t}\right)$ just as ruthlessly as it compresses wages. This mathematically describes the "intangible transition" of the digital era documented in Chapter 4: In a world dominated by software and platforms, value shifted away from physical assets (factories, machines) and toward unmeasured **intangible capital**—proprietary datasets, algorithms, and IP royalties. In the national accounts, returns to these intangibles often show up as "residual profits" rather than payments to traditional capital equipment (see Brynjolfsson et al. 2021).

However, the AI era may force a pivot yet again. Unlike the "capital-light" software startups of the 2010s, modern AI is aggressively physical. It requires massive tangible

---

competitive framework, the factor shares sum to only $\frac{1}{\mu}_t$, leaving a residual. As shown in the third term of $Eq.25$, this residual $\left(1 - \frac{1}{\mu}_t\right)$ is exactly captured by the Profit Share, ensuring that the total accounting identity holds.

investment in energy-intensive data centers, custom silicon fabrication, and grid infrastructure (see Chapter 10). This creates a new tension in the model:

- Will the sheer scale of this physical build-out force the capital share (α) back up?
- Or will the scarcity of the *intangible* design (the foundation model itself) continue to drive μ_t higher, keeping the "Rent Share" dominant even as the physical footprint of the industry explodes?

This model does not pre-judge that outcome; it simply gives us the "dashboard" (via α vs μ_t) to track it in real time.

Bringing Diffusion back in: how AI affects profits

I can now investigate AI's influence on profits through the two frictions introduced in the model; continuing with the ongoing thread, μ_t , and, revisiting diffusion dynamics, λ_t . To do so, I combine the markup structure I've outlined above with my diffusion-based productivity structure; specifically: $Y_t = A_t^{eff} K_t^{\alpha} L_t^{1-\alpha} = \lambda_t A_t^{frontier} K_t^{\alpha} L_t^{1-\alpha}$.

Substituting into $Eq.(24)$:

$$\Pi_t = \left(1-\frac{1}{\mu_t}\right) \lambda_t A_t^{frontier} K_t^{\alpha} L_t^{1-\alpha} \tag{26}$$

I can draw two conclusions from this equation:

- **Diffusion (λ_t) expands the size of the pie** (more of the frontier gets embedded in production).
- **Markups (μ_t) tilt how the pie is split** between labor, capital, and profits.

The Gatekeeper Dilemma

The upshot is that there is a potential "Gatekeeper" Dilemma: While the $Eq.26$ is a static accounting identity, the history of the digital age suggests there's a dynamic constraint on it.

$Eq.(26)$ treats the markup (μ_t) and diffusion (λ_t) as independent parameters; microeconomic theory dictates they are endogenous to one another. Because $P_t = \mu_t MC_t$, a higher markup mathematically forces a higher price. And since demand for General Purpose Technologies is historically price-elastic ($\varepsilon > 1$), this higher price creates a drag on adoption.[275] Therefore, while the equation does not explicitly contain a term like $\lambda_t = \mu_t^{-\varepsilon}$, the dilemma arises from this implicit dependency: a firm that raises μ_t to maximize its slice of the pie will inadvertently suppress λ_t, shrinking its overall size.

Is the "Gatekeeper Dilemma" inevitable, however? Basic calculus suggests it is not. In fact, if AI firms can solve their coordination problem and embrace a "High Volume/Low Margin" strategy collectively, they can actually earn *higher* total profits.[276]

---

[275] Empirical research confirms that the explosive diffusion of PCs and the internet was driven by price declines of 25–30% annually, revealing that demand for general-purpose computing is highly price elastic (see Nordhaus 2007; Byrne and Corrado 2017).

[276] This coordination failure is analogous to a “Double Marginalization” problem. Because the AI value chain is modular—requiring chips, cloud infrastructure, foundational models, and application layers to function together—the final price facing the adopter is the aggregate of multiple independent markups. If each firm sets its markup (μ_i) in isolation to maximize its own profit (the “Defect” strategy in a Prisoner’s Dilemma), the resulting collective price will be inefficiently high, stifling the very diffusion (λ_t) they all rely on. No single firm has the unilateral incentive to cut margins to the optimal level ($\mu \approx 1$) because they would bear the full revenue loss while the ecosystem shares the benefit of the induced volume. Thus, the “High Volume”

To see this, I must move beyond the static identity represented by (24) and introduce a simple product-market constraint: in this subsection, I treat ***Y*as the equilibrium quantity sold**, determined by demand at the price implied by the markup. Up to now Y has been supply capacity from the production function; here I'm temporarily shifting to the product market and treating equilibrium output as demand-determined, with the firm scaling inputs to meet demand. Because price moves with the markup, the relevant objective here is nominal profit (dollar profit), not "real profit per unit of output."

I start with a constant-elasticity demand curve (see Varian 2014):[277]

$$Y(P_t) \;=\; Y_0\, P_t^{-\varepsilon}, \varepsilon > 1 \qquad (27)$$

Where $Y_0$is a positive constant capturing baseline market size (and all the non-price shifters of demand).

Next, price is a markup over marginal cost:

$$P_t \;=\; \mu_t\, MC_t \qquad (28)$$

---

equilibrium requires a binding coordination mechanism—such as the SSOs discussed in Chapters 2 and 6—to enforce a simultaneous reduction in markups across the stack.

[277] I acknowledge that assuming constant elasticity is a simplification for illustrative purposes. Real-world AI markets often feature network effects and high switching costs, which can create "winner-take-most" dynamics that complicate the elasticity parameter. However, the core intuition remains robust: even with network effects, a General Purpose Technology relies on broad adoption to generate value, meaning the penalty for restricting quantity via high markups remains severe.

Holding MC_t fixed for this local pricing exercise (the point here is how pricing power interacts with diffusion), nominal profit is:

$$\Pi_t^N = (P_t - MC_t)\, Y(P_t) \tag{29}$$

Substituting $P_t = \mu_t MC_t$ and $Y(P_t) = Y_0 P_t^{-\varepsilon}$ yields a clean reduced form:

$$\Pi_t^N(\mu_t) = (\mu_t - 1)MC_t \cdot Y_0(\mu_t MC_t)^{-\varepsilon} = \underbrace{Y_0\, MC_t^{1-\varepsilon}}_{\text{constant } C>0}\, (\mu_t - 1)\mu_t^{-\varepsilon}$$

I differentiate with respect to μ_t. Since the constant $C \equiv Y_0 MC_t^{1-\varepsilon} > 0$ does not affect the optimum:

$$\frac{d\Pi_t^N}{d\mu_t} = C\, \mu_t^{-\varepsilon-1}(\varepsilon - (\varepsilon - 1)\mu_t)$$

Setting the derivative to zero gives the familiar Lerner monopoly condition I outlined in Chapter 4, $Eq.\,4.2$:

$$\mu^* = \frac{\varepsilon}{\varepsilon - 1}, \varepsilon > 1$$

Equivalently, the Lerner index, which I outlined in Chapter 4, $Eq.\,4.1$, is:

$$\frac{P_t - MC_t}{P_t} = 1 - \frac{1}{\mu^*} = \frac{1}{\varepsilon}$$

Now the key comparative-statics point that defines the Gatekeeper's Dilemma falls directly out of the derivative sign:

- If $\mu_t > \mu^*$, then $\frac{d\Pi_t^N}{d\mu_t} < 0$, so **lowering** the markup **raises** total profit.
- If $\mu_t < \mu^*$, then $\frac{d\Pi_t^N}{d\mu_t} > 0$, so **raising** the markup raises profit (up to μ^*).

I draw an important conclusion from this condition:

- If the current markup μ_t is set too high relative to the elasticity of demand, the firm is destroying its own value. In highly elastic markets (large ε), the profit-maximizing markup $\mu^* = \frac{\varepsilon}{\varepsilon-1}$ sits close to 1. In that case, a "high-margin/low-adoption" strategy is self-defeating: firms lose margin per unit when they cut markups, but the demand expansion is so large that **total profits rise**, while diffusion accelerates.

The upshot is that, in the AI era, "greed" may not look like the caricature of the monopolist hoarding a scarce resource to jack up prices. That is the strategy of a boutique luxury good, not a General Purpose Technology. Instead, because the demand for foundational AI is highly elastic—meaning its value comes from being everywhere, not just somewhere—the "Gatekeeper" who tries to maintain high margins is effectively lighting money on fire. The math shows that any attempt to price above the optimal markup ($\mu*$) destroys the volume of the market (Y) faster than it increases the profit per unit ($P - MC$).

Commodification Can Also Drive Profits Down

Of course, there is always the possibility that **temporary AI rents fade as diffusion rises.** While early movers may earn temporary rents, these may become compressed as AI becomes standardized, commodified, and widely available. Formally:

$$\mu_t = \bar{\mu} + \eta(1 - \lambda_t), \bar{\mu} \geq 1, \eta \geq 0 \quad (30)$$

Where η is a **sensitivity parameter capturing the "scarcity premium" or "novelty rent" associated with early-stage technology.** It measures how much pricing power firms possess simply because the technology is new, rare, and lacks standardized substitutes.

I draw two conclusions from this equation:

1. Early on (λ_t small), markups/rents are high (frontier firms have advantages).

2. As diffusion rises ($\lambda_t \to 1$), the "AI-specific" portion of markups fades and $\mu_t$returns toward a baseline $\bar{\mu}$.

There are several implications that flow from this logic to consider:

1. $Eq.30$ explains why the Gatekeeper approach is rational behavior for firms in the short run. Since firms know that margins will inevitably compress $as\ \lambda_t \to 1$ (due to commodification), they face an incentive to engage in **intertemporal price skimming**: extracting maximum rents (η) early while λ_t is low, effectively "harvesting" the market before it becomes competitive.
2. If $Eq.(30)$ holds, the "Infrastructure Layer" (chips and base models) will eventually become a low-margin utility ($\mu \approx \bar{\mu}$). This implies that long-run profitability must migrate **"up the stack"** to the Application Layer or Services Layer, where differentiation is harder to erode. This predicts a shift in the AI business model from selling "compute" (commodity) to selling "solutions" (differentiated).
3. However, there is always the opposing possibility that platforms with network effects could make μ_t *rise* with diffusion instead, entrenching a winner-take-most dynamic.

Therefore, $Eq.(30)$ serves as a mathematical litmus test for regulators, grounded in the logic of the Lerner Index, which I discussed in Chapter 4 and 7 (see Lerner 1934). In a functioning market, markups must compress as the technology matures. If we observe diffusion (λ_t) reaching saturation levels (e.g., 0.8 or 0.9, implying 80–90% adoption) but the markup (μ_t) remains stuck at high levels—defying the commodification pressures that typically force P toward MC—it suggests that the "Scarcity Premium" (η) has been replaced by **"Monopoly Entrenchment"** (artificial barriers to entry).

The Link Between Markups and the "Scale Effect"

The threat of Monopoly Entrenchment creates a final, crucial bridge between the "Profit" dynamics I just modeled and the "Labor" dynamics I am about to explore. As I will detail in the next section, a primary mechanism that saves jobs during automation waves is the **Scale Effect**: the idea that efficiency lowers prices, which boosts demand, which requires more workers to fulfill.

However, $Eq.(30)$ reveals a short-circuit. If the "Markup Wedge" (μ_t) remains high, the efficiency gains from AI ($\lambda_t A_t^{\text{frontier}}$) are captured as profits rather than passed through as lower prices. If prices do not fall, the expected demand surge never happens.

Consequently, the "Gatekeeper" scenario is doubly damaging: it not only siphons value from the economy in the form of rents, but it effectively blocks the "Scale Effect" from ever materializing. In this scenario, workers suffer the **Substitution Effect** (being replaced by AI) without enjoying the offsetting **Scale Effect** (more volume). The fate of the AI workforce, therefore, is inextricably linked to the pricing power of the AI firm. The impact of AI on jobs is the topic I turn to now.

AI AND JOBS

On the one hand, some research on AI and jobs is quite bearish. Chopra et al.'s (2025) so-called Iceberg Index estimates that current AI systems can technically perform skills corresponding to roughly 11.7% of U.S. wage value (about $1.2 trillion), far exceeding the "visible" exposure concentrated in computing (about 2.2%), and spanning hundreds of occupations. Bridging this theoretical exposure to actual market outcomes, Hui et al. (2024) provide micro-evidence from the gig economy: shortly after the release of major generative AI tools, freelancers in high-exposure fields like writing and coding saw a 2% drop in monthly contracts and a 5.2% decline in earnings, with highly skilled workers often suffering steeper drops than their less experienced

peers. Macro-wise, Ozkan and Sullivan (2025) report that unemployment increases from 2022–2025 are larger in more AI-exposed occupations, albeit they do not argue that this is a causal relationship.

On the other hand, there was compelling evidence that AI created more jobs than it destroyed by boosting workers' output and enabling new, specialized roles. Using occupation-level measures of labor demand and task requirements, Chen et al. (2024) report that generative AI is associated with declining demand and reduced skill requirements in more structured cognitive roles, but rising demand and greater skill complexity in jobs organized around human–AI collaboration.

Finally, a 2025 report analyzing usage patterns on Claude.ai following the deployment of Claude 3.7 Sonnet—an LLM featuring enhanced agentic coding capabilities and an "extended thinking" mode—shows that AI primarily enhances human capabilities, rather than replacing them (see Anthropic 2025e). Drawing on anonymized conversation data from Claude.ai users, the report documents that augmentation significantly dominated automation, with the former comprising 57% of all Claude usage versus 43% for the latter.[278] The absence of any occupation

[278] Augmentation is comprised of three types of activities: Learning, where users ask Claude for information or explanations, represented 28.0%, and was the largest single category; Task Iteration, where users iterate and refine outputs with Claude, accounted for 24.8%; Validation, used for checking or verifying outputs, accounted for just 4.1%. Among task iteration leaders in the augmentation category, Copy Writers lead at 57.7%, primarily writing advertising copy for publication or broadcast. Editors follow at 54.8%, preparing, rewriting, and editing copy, while Instructional Designers show 44.1% task iteration, developing instructional materials and lesson

category where automation exceeded augmentation suggests that, at least in 2025, current AI technology functioned best as a human enhancement tool rather than a human replacement.[279]

Unsurprisingly, occupations involving direct human interaction such as education, counseling, and healthcare showed the highest augmentation rates; writers and editors evidenced the highest task iteration rates, using Claude as a collaborative partner in the creative process rather than a content generator.[280] However, augmentation prevailed even in highly technical

---

plans. Meanwhile, automation is comprised of two types of activities: Directive interactions, involving direct task completion with minimal human involvement, accounted for 30.4%; Feedback Loop, which involves iterative automation processes, represented 12.7%. Among feedback loop leaders in the automation category, Network and Computer Systems Administrators top the list at 44.9%, diagnosing and resolving technical problems. Web Administrators follow at 41.0%, determining sources of web problems and taking corrective action, while Software Developers show 30.5%, primarily modifying existing software to correct errors.

[279] The most balanced occupations, approaching a 50-50 split between augmentation and automation, include Production occupations at approximately 51% augmentation, Protective Services at around 55% augmentation, and Computer and Mathematical occupations, also at about 55% augmentation.

[280] The highest augmentation occupations include Community and Social Service roles at approximately 75% augmentation, which includes education and counseling roles. Healthcare Practitioners show around 72% augmentation, Sales and Related occupations demonstrate about

fields like software development and IT, where professionals used Claude to enhance their capabilities rather than replace their expertise.

This evidence is consistent with several resonant examples of firms that tried to replace their workforce with AI but failed. Consider several organizations that retrenched from earlier plans to swap people for AI. The clearest emblem was Klarna (Marcus 2025c). After boasting in 2024 that its OpenAI-powered assistant could handle two-thirds of support chats—"the equivalent work of 700 full-time agents" (Klarna 2024)—the company shifted in mid-2025 to hire back humans so customers could "always speak to a real person," launching a recruitment drive and re-emphasizing human service alongside the bot (Bloomberg 2025). In August 2025, the Commonwealth Bank of Australia scrapped a plan to replace 45 call-center employees with an AI voice bot after call volumes rose and service degraded; the bank apologized and offered to reinstate or redeploy the affected staff (Chalmers 2025). And a 2025 survey of 1,000 senior business leaders reports that, while 39% trimmed their workforce in the face of AI-related redundancies, 55% say those cuts were mistakes, inducing them to reverse course and instead invest in reskilling and more flexible hiring model that blends permanent employees with specialized contractors and digital agents (Orgvue 2025).

Adjudicating the AI Jobs Debate

What does this preliminary and mixed evidence tell us about whether AI will create more jobs or destroy more of them? To make better sense of it, it behooves me to lean on theory once again to help reduce the noise and amplify the signal.

70% augmentation, and Life, Physical, and Social Science fields show roughly 68% augmentation.

The natural starting point are "task-based" models (e.g., those influenced by Acemoglu and Restrepo 2019, such as Autor et al. 2003). They offer a convenient mathematical lens for seeing how AI might replace or augment individual tasks, especially routine ones. Conceptually, these models imagine a "job" as the sum of discrete tasks, each of which can be assigned either to labor or to capital (AI). Formally, we often see an additive aggregator: $J = T_1 + T_2 + \cdots + T_n$, where each task T_n can be reallocated from humans to machines. Once a task is automated, the labor share for that task diminishes, while AI's share increases.

Looked at through this lens, routinized jobs are at risk. White-collar workers—translators, entry-level coders, concept artists—face displacement as AI generates text and code cheaply. Similarly, industrial and warehouse jobs, once insulated by limited robotic dexterity, become increasingly vulnerable to replacement as AI-driven vision and planning systems integrate with automated hardware.

However, while the task-based approach is elegant, it can be misleading in modern economies where intangible capital plays a large role (see Chapter 4). By assuming labor is merely a sum of discrete inputs, it fails to capture value added that is more synthetic and based on tacit contributions. In many real-world settings, a "job" isn't simply a bundle of tasks that can be unbundled one-by-one without consequence. Instead, a job title calls on the employee to engage in multiple, interacting responsibilities: preserving the capital stock, coordinating projects across teams, mentoring subordinates, maintaining the firm's reputation, and continuously updating intangible capital such as proprietary methods, institutional knowledge, or organizational culture (see Alston et al. 2018). These interactions can have a multiplicative (rather than additive) effect on how work is performed and any individual task cannot be sliced

out and automated with minimal repercussions on the rest of the job (Bresnahan et al. 2002; Milgrom and Roberts 1990).

Consequently, while some researchers attempt to incorporate "complementarities" or "task bundles" to address these feedback effects, the resulting models become significantly more complex and can lose the simplicity that makes the task-based approach so appealing in the first place (Acemoglu and Autor 2011; Autor and Handel 2013). For example, it is very hard to apply this framework to childcare work, law enforcement, wildlife management, fashion modeling, religious ministry, stonemasonry, and flight attending—all involve social interaction, subtle human judgements, and complex physical tasks.

As AI spreads in economies that are already rich in intangible capital where jobs rely heavily on cooperation, creativity, and the preservation of firm-specific knowledge, the assumption of additive job tasks may be wholly inadequate (see Mavens Global 2025). Real workplace roles in these contexts involve interdependent responsibilities and relationship-building. This means it is impossible to posit that "AI automates xx% of tasks and labor's share decreases proportionally."

Labor Supply and Demand Framework for AI's Effects on Jobs

I can blend these criticisms of task-based approaches to labor demand with the tentative evidence about AI's lack of job destruction outlined above to inform a simple partial-equilibrium model for a single sector/industry. While it retains some of the task-based approach to employment's core elements, it transcends its limitations by aggregating these complex, multiplicative interactions into a net shift of the labor demand curve. This allows us to model how the productivity-driven "scale effect"—fueled by intangible capital—can overpower the

"displacement effect," increasing the total value of a job even as specific tasks within it are automated.

This model is represented by Figure 11.1. The wage w is mapped on the vertical axis and the quantity of labor L lies on the horizontal axis. In turn, labor demand is represented by a downward-sloping curve $D(w)$—relating labor demanded L_D to the wage w—and labor supply is given by an upward-sloping (or vertical) curve $S(w)$, reflecting worker availability at various wage rates. The original labor-demand curve D_0 may shift to D_1, either rightward if the scale effect dominates, or leftward if automation overwhelms the scale effect.[281]

The ultimate effect of AI on jobs revealed by this model is that it depends on which force dominates between substitution and scale forces. On the one hand, AI reduces the cost of specific tasks through within-firm substitution. If final demand is inelastic, employment might fall as substitution reduces total labor usage. If AI is strongly complementary to human labor and final

[281] Technically, the demand curve for labor may repeatedly shift outward over time, tracing an upward-sloping path for labor demand over the long run. As a society gets richer, and as innovation intensifies, there will be increased demand for goods and services (hence, demand curves will keep shifting out) and, in turn, increased demand for the domestic labor who make these goods and services. In other words, employers will have an increased willingness to pay laborers—especially because innovation will make them more productive. While the scholars who model these dynamic effects focus on skilled (educated) labor to explain why the returns to college degrees increased even as the supply of graduates rose (Acemoglu 2002; Acemoglu 2009; Goldin and Katz 2008), conceivably the same process may apply to unskilled labor too (see Bessen 2015). More on this below.

demand is elastic, the new equilibrium may feature higher employment. On the other hand, AI may allow firms to offer better or cheaper products/services, possibly expanding production scale if demand is elastic—AI may make products or services cheaper and better (shifting the supply curve). Despite task-level automation, labor demand for tasks AI cannot do alone or tasks that AI enhances should rise in this case. In short, whether net employment rises or falls depends both on whether AI's cost-saving automation overshadows or gets overshadowed by product-market growth and labor demand elasticity.

The Macro-Micro Synthesis: The Condition for Job Creation

Another way to conceptualize this dynamic is to briefly return to the macroeconomic model I explored above. There, I explored **scale effects** in the form of lower prices/higher quality expanding output. And I also explained that whether scale effects materialize depends on **passthrough**: do productivity gains reduce prices, or are they captured as markups/profits?

However, the magnitude of the markup wedge is not just about profits; it determines the fate of the worker. I can formally link the "Labor Demand Shift" in Figure 11.1 back to the "Markup Wedge" (μ_t) using the pricing condition derived earlier.

Figure 11.1 A Partial-equilibrium Labor-market Model for a Single Sector.

[INSERT FIGURE HERE]

Notes: This labor market model illustrates AI's dual impact on employment through competing forces. The original equilibrium occurs where labor demand D_0 intersects supply S. When AI enters an industry, labor demand may shift to D_1 (rightward shift) if the scale effect dominates—meaning AI's efficiency gains expand production sufficiently to increase overall labor demand for complementary tasks. Conversely, demand shifts to D_2 (leftward shift) if the substitution effect dominates—when AI automates enough tasks that displacement exceeds any production

expansion. The net employment outcome depends critically on two factors: (1) the elasticity of final product demand—elastic demand favors employment growth as lower costs drive market expansion, while inelastic demand limits this offsetting mechanism; and (2) the degree of task complementarity versus substitutability between AI and human workers. The upward-sloping supply curve S remains stable, reflecting worker availability at different wage levels. The resulting equilibrium wage and employment level reveal whether AI acts primarily as a complement (raising both) or substitute (reducing both) for human labor in the specific industry context.

Since price is a markup over marginal cost ($P_t = \mu_t \cdot MC_t$), and marginal cost falls as effective technology rises ($MC_t \propto 1/A_t^{\text{eff}}$), I can define the **"Transmission Mechanism"** for the scale effect:

$$P_t = \mu_t \cdot MC_t, MC_t \propto \frac{1}{A_t^{eff}} = \frac{1}{\lambda_t A_t^{frontier}} \Rightarrow P_t \propto \frac{\mu_t}{\lambda_t A_t^{frontier}}$$

(31)

This equation provides the **"tipping point"** for the labor market. It tells us exactly what forces are fighting for the soul of the Labor Demand curve:[282]

- **The Numerator (The Villain):** The Markup Wedge (μ_t). If this rises, prices stay high, blocking the scale effect.

[282] Readers should note a deliberate simplification here. In a full general equilibrium model, a rise in the markup (μ_t) would depress factor prices (w_t, r_t) via $Eq.21$, which would theoretically lower nominal marginal cost (MC_t). To avoid circularity and keep the intuition sharp, $Eq.31$ focuses on the partial equilibrium transmission: it isolates how the markup wedge acts as a multiplier on price for any given level of underlying technology and factor costs.

- **The Denominator (The Hero):** The Diffusion of Frontier Tech ($\lambda_t A_t^{\text{frontier}}$). If this rises, costs fall, driving prices down and volume up.

Therefore, I can draw two powerful conclusions that define the "Condition for Job Creation":

1. **The Scale Effect Dominates (Labor Demand Shifts Right):** If the efficiency gains in the denominator ($\lambda_t A_t^{\text{frontier}}$) grow **faster** than the markup in the numerator (μ_t), then prices (P_t) fall meaningfully. This triggers the massive expansion in output required to soak up displaced labor and creates new complementary roles.
2. **The Substitution Effect Dominates (Labor Demand Shifts Left):** If the "Gatekeepers" raise markups (μ_t) roughly in sync with technological improvements, prices (P_t) remain flat. The scale effect is short-circuited. In this scenario, the economy suffers the displacement of automation without enjoying the expansion of volume.

The upshot is that profits and market power are tied directly back to jobs. High markups do not just hurt consumers; they actively destroy the specific economic mechanism—the **Scale Effect**—that protects workers from obsolescence.

Making Sense of What AI Complementarities Mean

However, securing the macroeconomic conditions for job growth—keeping markups low to fuel the scale effect—is only half the equation. The other half depends on the nature of the work itself. Even if the “Gatekeepers” allow output to expand, humans will only share in that growth if they offer a capability distinct from the machine.

If AI is a perfect substitute for human effort, a rightward shift in demand will simply lead to more servers, not more workers. Therefore, the “Rightward Shift” in Figure 11.1 relies entirely on the persistence of **Comparative Advantage**.

How might this play out? Consider that AI often has absolute and comparative advantages in speed, memory, and data processing, while humans retain absolute and comparative advantages in non-codifiable expertise, domain knowledge, and interpersonal skills. his interplay yields gains from trade, governed by the principle of Comparative Advantage (Ricardo 1817). It is vital to define this precisely, as it is often confused with Absolute Advantage. Even in the case where an AI model possesses an Absolute Advantage in every single task—meaning it can generate code, write emails, and diagnose patients faster and more accurately than any human—it does not follow that humans become obsolete. Economic efficiency dictates that resources should be allocated where their relative advantage is highest. Formally, a human retains a comparative advantage in $Task\ A$ (e.g., Strategic Oversight) over $Task\ B$ (e.g., Data Processing) if the ratio of their productivity (A) to the AI's productivity is higher in Task A than in Task B:

$$\frac{A^{human}(Task\ A)}{A^{\mathrm{AI}}(Task\ A)} > \frac{A^{human}(Task\ B)}{A^{\mathrm{AI}}(Task\ B)} \tag{32}$$

Where:

- $A^{human}(Task\ X)$ is the productivity (output per unit of time) of a human performing a specific task.
- $A^{\mathrm{AI}}(Task\ X)$ is the productivity of the AI system performing the same task.

The inequality states that a human retains employment in $\boldsymbol{Task\ A}$ not because they are "better" than the AI (indeed, A^{AI} could be higher in both numerators), but because the **ratio** of their performance is superior. As long as the gap between human and machine capabilities is smaller in $Task\ A$ than in $Task\ B$, economic efficiency dictates that the AI should be allocated to $Task\ B$ (where its leverage is greatest), leaving $Task\ A$ to the human.

Thus, integrating AI into enterprises becomes an exercise in complementary specialization rather than simple substitution. Society can maximize total economic output only if AI handles repetitive or data-heavy tasks, such as summaries and information retrieval, and humans can focus on conceptual insight, ethical reasoning, or interpersonal interactions.

Integrating AI into enterprises or working alongside it calls on human oversight, creativity, dexterity, or nuanced judgment. For example, in construction or engineering, AI might excel at calculations and modeling, but human understanding of fundamental principles remains essential for interpreting and supervising AI output. Similarly, in medical diagnosis, while AI can process vast data efficiently, physician expertise in anatomy, pathology, and clinical context remains crucial for complex cases and ethical decisions. Indeed, David Autor (2015), citing Kremer (1993), famously observed that "O-Ring" tasks—those requiring high-stakes accountability and synthesis—remain human domains. The AI may be "smarter," but its computational power is too valuable to waste on oversight when it could be solving protein folding.

This is also consistent with work by Ethan Mollick, who argues that the most effective use of AI is not as an autonomous replacement for human workers, but as a collaborative partner in a "cyborg" model, where the combined output of the human-AI team is greater than the sum of its parts (Mollick 2024). In this framework of "co-intelligence," AI acts as a tireless assistant, handling routine analytical and generative tasks, which frees up human workers to focus on creativity, strategic judgment, and empathy (see Mollick n.d.).

Drawing on the model represented by Figure 11.1, what are some new jobs or changing roles within extant job categories that AI may engender within specific sectors/industries when a scale effect dominates a substitution effect? Below I consider several examples and elucidate the

reasons why more jobs may be created than destroyed across diverse industries in the wake of the AI revolution.

First, consider the AI stack itself. New roles in AI development, maintenance, and oversight may continue to emerge—specifically, more jobs for data scientists, MLOps engineers, AI prompt engineers, and domain specialists who finetune models for specific industries. Here, a rightward shift to D_1 occurs because AI development itself represents a new product market with highly elastic demand, where improved AI capabilities generate exponentially more applications and require proportionally more human oversight.

Take IBM.[283] To support its massive push into the "AI Stack" with its Watsonx platform, the company had to aggressively hire for roles that didn't exist a decade ago—AI ethics governors, data curators for Large Language Models, and full-stack AI architects. While the company simultaneously automated routine internal HR workflows, the explosive demand for its new AI products meant that overall headcount in these technical divisions rose. Here, the "Scale Effect" is visible: the invention of a new product (enterprise AI) shifted the demand curve for specialized labor rightward, overwhelming the substitution effects in its back office.

Second, in some industries workers may specialize in writing precise instructions in natural language to squeeze out efficiency gains from AI at both the front end and back end. In this scenario, AI generates initial versions of documents or solutions with human help while experts subsequently review, adjust, and approve these products. Similarly, enterprises may want to guarantee that every automated decision remains traceable and fixable (auditable and debuggable), calling on them to develop new governance features such as records of AI-

[283] This paragraph draws on Lin (2025).

generated code changes, systems to track and manage changes to code, and rollback mechanisms that can revert to previous versions. These quality assurance and governance roles exemplify how the scale effect dominates when AI creates complementary tasks that didn't previously exist, shifting labor demand rightward even as some routine tasks are automated. When AI dramatically reduces production costs through automation, highly elastic demand for certain products/services drives a scale effect that overwhelms labor substitution.

Consider software engineers. As AI takes on the repetitive "busywork" of coding—auto-generating boilerplate, debugging straightforward errors, even scaffolding entire modules—it paradoxically elevates the value of the very computer-science foundations it automates (Partovi 2025). This shifts the engineer's marginal product curve upward/rightward. By lowering barriers to entry, AI tools enable more junior developers to contribute productively sooner, expanding the pool of coders and creating demand for new roles. Among these, orchestration may become a crucial, high-level task—rather than writing individual lines of code, a software engineer may, akin to a conductor leading an orchestra, ensure all the different AI tools and software pieces work together smoothly. This may mean designing reliable ways for AI-powered systems to handle unexpected problems and building easy-to-use controls so people can guide and fix these smart systems (Dabbir 2025; Daws 2025).

Finally, entirely new business models may emerge, leading to the creation of new service providers or startups and more jobs.

Take driverless cars. As automobiles embrace full autonomy, they will likely evolve from standalone products into on-demand, experience-driven platforms. If a car becomes a rolling office or wellness pod, entirely new employment categories emerge to service that platform. Designers will migrate from hospitality to automotive UX; fleet planners will manage dynamic

pricing for “mobility-as-a-service”; and remote monitoring specialists will act as air-traffic controllers for ground transit. In this scenario, the “Scale Effect” dominates because the technology doesn't just replace drivers—it lowers the cost of transport so drastically that it creates a massive new market for mobile services, shifting the labor demand curve decisively rightward to D_1 (see Figure 11.1). Hence, total job gains would far exceed the jobs lost to driving automation.

MODELING EXERCISE’s LESSONS

The formal exercises in this chapter clarify that the economic impact of AI is not determined solely by how smart the models get (A^{frontier}), but by the structural constraints that govern their deployment and pricing. By moving beyond simple "robots take jobs" narratives, the framework reveals that we are actually witnessing a race between three variable forces: Diffusion, Markups, and Comparative Advantage.

The core insight is that "Efficiency" does not automatically equal "Prosperity." If the efficiency gains from AI are trapped by low diffusion (λ_t) or captured by high markups (μ_t), the macro-level result is stagnation and inequality ("Vampire Capitalism"), even if the micro-level technology is miraculous. Conversely, the "Golden Age" scenario requires a specific alignment of incentives: firms must coordinate to lower markups to unlock the "Scale Effect," thereby transforming AI from a scarce luxury into a general-purpose utility that drives demand for human labor.

Table 11.1. Summarizing the Theoretical Framework’s Key Insights

Economic Concept	***Formal Condition/ Mechanism***	***Key Lesson for the AI Era***

The Standardization Gap	$Aeff = \lambda_t \cdot A^{frontier}$	**Potential ≠ Production.** Massive investment in chips and models ($A^{frontier}$) yields minimal productivity growth until standardization (λ_t) reduces the friction of adoption.
The Gatekeeper's Dilemma	Profit rises if $\mu_t \downarrow$ when $\epsilon > 1$	**Price Cuts maximize Profit.** In elastic markets (like general compute), keeping markups high ("skimming") destroys value. Collaboration to lower costs is the profit-maximizing strategy long-term.
The "Vampire" Wedge	$(w_t) = \frac{(1-\alpha)}{\mu_t}\left(\frac{Y_t}{L_t}\right)$	**Decoupling is Structural.** If markups

		(μ_t) rise alongside productivity $\left(\frac{Y_t}{L_t}\right)$, real wages ($w_t$) will mathematically flatline. Inequality is a market structure problem, not just a skill problem.
The "Scale Effect" Condition	$P_t \propto \frac{\mu_t}{\lambda_t A^{frontier}}$	**Markups Kill Jobs.** Labor demand only shifts right (job creation) if efficiency gains (λA) overpower markups (μ), causing prices (P_t) to fall and output to surge.
Comparative Advantage	$\frac{A^{human}(Task\ A)}{A^{AI}(Task\ A)} > \frac{A^{human}(Task\ B)}{A^{AI}(Task\ B)}$	**You Don't Have to be "Smarter."** Humans retain employment not by outperforming AI, but by specializing in tasks

		where the opportunity cost of using AI is too high (e.g., judgment, oversight).

Notes: λ_t represents the diffusion rate (0 to 1); μ_t is the markup wedge ($\frac{P}{MC}$); ε is the price elasticity of demand; α is the capital share parameter in the Cobb-Douglas function; A^{frontier} is the theoretical maximum capability of the technology; $\frac{Y_t}{L_t}$ is labor productivity. For the "Vampire" Mechanic, the wage equation (w_t) demonstrates that even if productivity ($\frac{Y}{L}$) doubles, real wages will not rise proportionally if the markup (μ_t) simultaneously increases, mathematically proving the mechanism of "decoupling." For the Scale Condition, the price equation (P_t) serves as the "transmission mechanism." It shows that the deflationary pressure of innovation ($\lambda_t A^{\text{frontier}}$) can be fully neutralized by the inflationary pressure of rising markups (μ_t), preventing the output expansion required to sustain employment. Regarding Comparative Advantage, the inequality condition holds even if $A_{AI} > A_{human}$ for both tasks. Employment depends on relative ratios, meaning humans remain economically essential in "O-Ring" tasks (e.g., oversight, judgment) where the opportunity cost of deploying AI is prohibitively high.

How this Chapter Fits with Earlier Chapters' Take on Superstar Firms

The current modelling exercise shows that any markup wedge (μ_t) mathematically reduces the share of output going to workers; yet readers will recall that in Chapters 4 and 5, I championed the "Superstar Firms" of the Third Industrial Revolution. These firms leveraged massive investments in intangible capital to generate foundational general-purpose technologies and, crucially, diffused them throughout the economy. By acting as innovation hubs and coordinating

via Standards-Setting Organizations (SSOs), they standardized best practices and allowed knowledge to flow to rivals and partners alike, thereby raising the productivity of the entire ecosystem.

How can these two realities coexist?

As Table 11.2 makes clear, the answer lies in the distinction between the Labor Share (the slice) and the Real Wage (the pie). In the "Superstar" era documented in Chapter 4, high markups were largely Ricardian Rents—rewards for extreme efficiency. While these high margins did indeed depress the labor share, contributing to the decoupling between productivity and wages that began in about 1980 (see De Loecker et al. 2020; Barkai 2020), the firms were so productive that they massively expanded the total economic pie ($\frac{Y_t}{L_t}$). For workers, the "Scale Effect" overpowered the "Markup Wedge"—they got a smaller percentage, but of a rapidly growing total, so their real wages rose.

The danger lies in the AI rocket falling back down to the ground before it reaches the escape velocity of full-fledged diffusion because standardization or organizational adaptation flounder. If "Gatekeepers" somehow manage to block diffusion ($\lambda_t \to 0$), markups remain high (suppressing the labor share), but without the compensating explosion in output. In this scenario, workers are hit twice: they suffer the extraction of the wedge without enjoying the expansion of the scale effect.

Table 11.2 The Vampire versus the Superstar

	Source of the Markup	*Impact on Labor Share (w/Productivity)*	*Impact on Real Wage Level*

The "Superstar" Creative Destruction Economy (Chapters 4 & 5)	*Efficiency (Ricardian)*: Low Marginal Costs allow high margins. High Innovation + High Diffusion.	The wedge widens. Workers capture a smaller slice of value.	The "Pie" (Productivity) grows so fast that it overpowers the wedge.
The "Vampire" Economy (Chapter 11)	*Pure Market Power (Monopoly)*: Lack of competition allows high prices. High Barriers + Low Diffusion.	The wedge widens. Workers capture a smaller slice of value.	The "Pie" does not grow fast enough to offset the extraction.

Notes: The table resolves the apparent paradox that high-markup firms can be beneficial in one era and destructive in another. In both scenarios, a high markup wedge (μ_t) mathematically depresses the Labor Share (the slice of the pie going to workers). The difference in the two scenarios lies in the Size of the Pie ($\frac{Y_t}{L_t}$). In the PC/Smartphone era (Chapter 4), high markups were typically Ricardian Rents constrained by Limit Pricing, pricing just below the cost of the inefficient fringe (Bain 1956; Milgrom and Roberts 1982). Because the technology diffused widely (λ_t was high), the competitive fringe forced the superstar to innovate relentlessly.

Consequently, the "Scale Effect" (massive productivity growth) overpowered the "Markup Wedge," causing real wages (w_t) to rise even as the labor share fell. In the "Gatekeeper" scenario modeled in this chapter, the markup becomes a Monopoly Rent. If diffusion is blocked ($\lambda_t \to 0$), the competitive fringe vanishes, and the "Scale Effect" stalls. Without the exploding pie of productivity to compensate, the high markup simply extracts value from a stagnant system, causing real wages to flatline.

The Modeling Strategy's Limitations and Extensions

The questions I have chosen to address in this chapter—such as why productivity gains lag frontier capability, how markups mediate distribution, and under what conditions the Scale Effect preserves employment—remain tractable because the model holds other margins fixed. While a framework that attempted to endogenize everything simultaneously would sacrifice the analytical clarity that makes these mechanisms visible, I leave it to future research to experiment with some of the extensions I outline below.

Endogenous Growth

While this book's historical focus has compelled me to stress that we cannot treat innovation as manna from heaven, but instead as the downstream economic output of institutions, policies, IP, R&D investments, and the disruption of incumbent firms by new entrants, the framework I introduce in this chapter takes the frontier as given. For good reason: The "binding constraint" on prosperity and other downstream outcomes circa 2025 was not the speed of innovation at the frontier, but the speed of AI absorption across the rest of the economy.

Because I chose to focus on the *adoption* of that technological frontier and its ramifications on economic efficiency and distribution, however, my modeling exercise remains distinct from endogenous growth models. Romer (1990) and Aghion and Howitt (1992), for

example, endogenize the *creation* of the frontier technology, modeling the R&D decisions and patent races that generate new innovations.

The upshot is that my framework cannot capture feedback loops that may prove decisive over the long run. For example, might a "High-Diffusion" equilibrium itself accelerate frontier progress by expanding the installed base of users generating training data? Would concentrating resources in well-capitalized incumbents paradoxically sustain frontier investment, ala Schumpeterian approaches to endogenous growth theory (See Acemoglu 2009). Practically-speaking, a self-reinforcing cycle where diffusion and frontier advancement feed into each other might meaningfully alter the model's long-run predictions, particularly regarding whether the labor share recovers after the diffusion phase saturates.

Skill Biased Technological Change

My framework also abstracts from skill-biased (and task-biased) technological change. In the formal model, labor enters as a single aggregate input L_t, so “the wage” is an economy-wide average tied to the marginal product of labor. That’s enough to reason about whether AI raises overall productivity, and whether the gains flow to labor, capital, or markups—but it cannot predict how those gains are divided *within* the labor force: whether the college premium rises, whether middle-skill work is hollowed out, or whether earnings inequality compresses because AI brings novices closer to experts—all pressing questions in the minds of many.

The modern baseline for analyzing these patterns—explaining why incentives often shape innovation to be biased against specific groups—is the literature on directed technical change (Acemoglu 2002) and the task-based framework that links these biases to occupational polarization (Acemoglu and Autor 2011). They mathematically parse how specific technologies differentially impact the marginal productivity of high- versus low-skill workers, thereby

explaining distinct phenomena like the rising college premium or the polarization of the job market.

For AI in particular, these inter-labor group distributional consequences remain genuinely ambiguous. Early evidence on generative-AI "copilot" tools suggests that assistance can deliver especially large productivity gains for less experienced workers within an occupation and can compress the performance distribution—effects that look *skill-equalizing* in the short run (Brynjolfsson et al. 2025; Noy and Zhang 2023). At the same time, AI could still prove skill-biased over the longer run as firms reorganize production, redesign jobs, and shift hiring toward workers who can supervise systems, integrate tools into workflows, and convert model output into high-stakes decisions. Autor (2024), for example, emphasizes a distinct possibility: that decision-support AI can *extend expertise* to a broader set of middle-skill workers—potentially counteracting some of the "hollowing out" dynamics of the prior IT era.

Incorporating this margin formally would require splitting labor into skill groups (e.g., high- and low-skill L_t^H, L_t^L) or adopting an explicit task-allocation structure. That extension is feasible, but it would substantially expand the machinery relative to the questions this chapter prioritizes.

Globalization Dynamics

My approach also abstracts from international trade in AI services and cross-border technology flows, even though, as I outlined in Chapters 4 and 6, elements of the AI supply chain, especially semiconductors, remained deeply globalized at the time of this writing. Indeed, models trained in the United States may be deployed in Europe, fine-tuned on data from Southeast Asia, and run on chips fabricated in Taiwan. This creates interdependencies that my single-economy treatment cannot capture. For instance, if one jurisdiction imposes stringent AI regulations that slow

domestic diffusion, firms may offshore deployment to more permissive environments, complicating the link between domestic policy and domestic productivity. Similarly, export controls on advanced semiconductors—see the next chapter—could constrain frontier capability in targeted economies in ways that interact with diffusion dynamics. Full treatment of these spillovers would require a multi-country extension incorporating trade in both AI services and the physical inputs to AI production.

Demand Side Economies of Scale

Network effects represent another significant omission. The model treats the markup wedge as a parameter that may rise or fall with diffusion, but it does not formally incorporate the demand-side increasing returns that characterize platform markets (see Chapter 7). In industries with strong network effects, early adoption by some users raises the value of adoption for others, potentially generating winner-take-all dynamics that could either accelerate diffusion (if a dominant platform lowers barriers to entry) or entrench gatekeepers (if switching costs lock users into proprietary ecosystems). The ambiguous relationship between network effects and competition—whereby the same forces that drive rapid adoption can also cement market power—deserves formal treatment that this framework does not provide.

Heterogeneity in Demand

Finally, the constant-elasticity demand assumption underlying the Gatekeeper's Dilemma analysis may not hold uniformly across AI applications. Enterprise software markets, in particular, often exhibit low short-run price elasticity due to contractual lock-in, integration costs, and organizational inertia, even when long-run elasticity is higher (Farell and Klemperer 2007). This implies that the self-correcting mechanism I describe—where excessive markups

destroy the volume required to sustain profits—may operate slowly or incompletely in market segments where customers cannot easily switch providers.

CONCLUSION

The modelling exercise conducted in this chapter offers a sobering but ultimately empowering conclusion about AI's effects. The "AI Future" is not deterministic: the "Robot Apocalypse" (permanent labor displacement) and "The Great Stagnation" (no sustained productivity boom; see Cowen 2011; Gordon 2016) are not inevitable features of the technology itself. These gloomy possibilities are instead symptoms of a "High-Friction, High-Markup" equilibrium that reflects choices made by both policymakers and CEOs. If standardization lags and "Gatekeepers" entrench a rent-seeking business model, the equations predict a world of dazzling technical demos and maybe even a pronounced productivity boom, but also stagnant wages, where the efficiency gains of AI are siphoned off before they ever reach the labor market.

However, the alternative path—the "High-Volume, High-Diffusion" equilibrium—is equally visible in the math. In this scenario, the industry solves its collective action problem, embracing the interoperability conferred by standardization and pursues modest, but steady margins comfortably exceeding the opportunity costs of capital. Indeed, firms at the cutting edge of the semiconductor, internet, and telecommunications revolutions reinvested retained earnings to continue to innovate, furnishing humanity with Moore's Law, the smartphone, and the gig economy (Hall 2002; Lazonick 2009; Galetovic et al. 2021). They did so because they competed in a paradigm of creative destruction where their profits were not guaranteed but instead depended on continuing to provide "the next big thing," whether it was a more advanced processor or 4G (see Chapters 5 and 6).

If such a goldilocks scenario also takes hold for AI, the "Scale Effect" is unleashed. Prices fall, demand explodes, and the comparative advantage of human workers shifts from rote cognition to high-level synthesis, oversight, and empathy.

The challenge of the coming decade, therefore, is not just to build smarter models, but to build the market structure that forces those models to serve the many rather than enrich the few. As with the Third Industrial Revolution, the "Invisible Hand" will not do this work for us; it requires the visible hand of coordination, standardization, and, where necessary, smart regulation, to ensure the revolution fully diffuses and reaches its potential.

That was the magic of the Creative Destruction Paradigm I outlined in Chapter 3: the US federal government specialized along its comparative advantage, helping to solve market failures such as the need to help coordinate profit-seeking firms around the common goal of creating and commodifying new technology markets when they faced perverse incentives to withdraw into their own fortresses and cashout their initial rounds of investment. And high-tech firms responded by specializing along their own comparative advantage, maximizing profits over the long run by providing affordable goods, services, and ecosystems with mass appeal. Ergo: PCs, the internet, smartphones, and the gig economy.

However, this future is far from guaranteed: The next chapter analyzes the divergent innovation policies of the US, China, and the EU in the wake of the decline of the CDP. While all three polities shifted toward mercantilism, protectionism, and state intervention, their approaches differed. The US blended elements of the CDP with an increasingly aggressive industrial policy. China pursued a state-directed, application-focused strategy. The EU lagged these efforts considerably, and its bid to catch up yielded mixed results. I contextualize these developments within broader deglobalization trends, illustrating the pitfalls of industrial policies based on neo-

mercantilist attempts by these polities to fractionalize AI stacks, foster national champions, and engineer rents through nationalistic regulations.

CHAPTER 12: THE FRAGMENTED REVOLUTION

By the end of 2025, the Fourth Industrial Revolution was being forged in a world defined by borders. Physical borders constrained semiconductor manufacturing and energy grids. Legal borders enforced data sovereignty and allocated liability. Ideological borders governed the purpose and control of machine intelligence.

This fragmentation marked the definitive end of the "One World" digital order that characterized the Third Industrial Revolution. For thirty years, the internet was governed by a rough consensus of borderless exchange, where capital, components, finished devices, data, and code flowed relatively freely across a single global supply chain. At the time of this writing, that era seemed over. The digital commons that provided the raw fuel for the AI boom was being enclosed, replaced by a series of competing, gated ecosystems.

This chapter examines how the three major economic blocs—the United States, China, and the European Union—arrived at distinct governance equilibria during the early 2020s, despite confronting the same underlying technology.[284] The conventional framing of these

[284] For a fuller treatment of the material condensed in this chapter, see Menaldo (2026: Section 4), where I situate the "fragmented revolution" in the broader geopolitical context of deglobalization, protectionism, and competing governance models; (ii) provides a detailed, comparative account of each bloc's emerging industrial-policy playbook (including compute and infrastructure buildout, export controls and allied technology packages, and efforts to shape international standards); (iii) surveys the major regulatory fault lines that will condition AI diffusion—liability and the AI "black box" problem, privacy and cross-border data flows, intellectual property, labor and education, and local resistance to AI's physical footprint; (iv)

differences emphasized ideology: the United States was described as permissive and innovation-oriented, Europe as precautionary and rights-focused, China as authoritarian but strategically coordinated.

These characterizations captured real differences, but they also obscured deeper patterns. All three blocs moved toward greater statism and technological mercantilism during this period, though from different starting points and through different mechanisms. What requires explanation is not merely how they differed, but why their trajectories converged on fragmentation rather than harmonization.

A final implication of fragmentation is that power does not reside only in the three major blocs. In a world of chokepoints, smaller technologically advanced economies can wield outsized influence by controlling critical bottleneck inputs. In the semiconductor equipment layer, for example, the global AI industry is effectively dependent on a single Dutch firm, ASML, which is the sole supplier of the EUV lithography tools required to manufacture leading-edge chips (see Chapter 6). This reality complicates the usual "U.S. vs. China vs. EU" framing. Despite increasing export controls and regulatory Balkanization, the AI race is mediated by narrow, hard-to-replicate industrial capabilities that may not bend to the Big Three's will.

In the United States, the model that had governed digital innovation since the 1980s—what earlier chapters termed the Creative Destruction Paradigm—did not disappear, but it mutated. By 2025, it had devolved into a messy, transactional form of patrimonialism, where the

---

speculates about how these competing approaches may fragment AI goods and services into region-specific variants; and (v) how private ordering (standards, audits, procurement, and insurance) may increasingly fill governance gaps when state capacity or coordination lags.

Creative Destruction Paradigm and market mechanisms persisted to a certain degree, but outcomes increasingly depended on proximity to political power. Technological commercialization continued in the face of uncertainty, as regulatory postures shifted between and sometimes *within* administrations, and enforcement varied according to political priorities.

China consolidated a model that can be termed State-Centric Managerialism. Unlike the American approach, which remained largely reactive and relied on ex post litigation to address harms, Beijing's model was fundamentally proactive and preventative. It treated AI not as a neutral market good but as a political instrument that required alignment with state goals before reaching the public. Where the United States struggled with a chaotic patchwork of state laws and federal inaction, China built a comprehensive, vertically integrated regulatory stack—from data labeling standards to algorithm filings—that enforced ideological conformity as a prerequisite for market entry. This approach traded openness for predictability, reducing uncertainty about permissible uses while narrowing the scope of acceptable experimentation.

The European Union pursued what became a Regulatory-First model. Driven by the aspiration to export its values of safety, privacy, and fundamental rights through comprehensive legislation, Brussels bet that trustworthy and safe AI would beat speed of execution. The AI Act established ex ante obligations on developers and deployers, imposing risk classification, conformity assessment, and transparency requirements before systems could reach the market. Through what observers termed the "Brussels Effect," European regulators sought to set global standards by making compliance with EU rules a condition of access to its large consumer market. Yet this regulatory ambition increasingly collided with anxiety, even panic, as Europe's own AI sector fell further behind American and Chinese competitors.

The chapter proceeds as follows. I first examine why AI posed a distinctive governance challenge, focusing on the gap between technological capability and regulatory capacity. The third through fifth sections analyze the American, Chinese, and European responses in turn, identifying how each configuration emerged from prior institutional arrangements and produced characteristic tradeoffs. The seventh section considers what these three cases reveal about the broader trajectory of AI governance, emphasizing convergence toward fragmentation rather than harmonization. The chapter and book conclude by examining the stakes of this fragmented landscape for the future of innovation and the fate of the Creative Destruction Paradigm.

AI'S STRUCTURAL CHALLENGE TO GOVERNANCE

The rapid commercialization of artificial intelligence in the early 2020s did not simply intensify familiar debates over privacy, labor displacement, or market concentration. It exposed a structural mismatch between technological capability and the institutional frameworks charged with governing it. The commercialization of AI outpaced the regulatory frameworks designed to govern it, generating pressure across legislatures, agencies, and executive offices to act—without equipping those institutions with adequate tools for effective intervention.

Three features of AI development made this mismatch especially acute.

First, AI systems posed distinctive challenges to accountability. Many modern AI systems, especially complex machine-learning and deep-learning models, make decisions in ways that are opaque even to their creators. Users can see the inputs and outputs, but they cannot trace the internal steps the system takes to reach a particular decision—why, for instance, an algorithm flagged one applicant as high-risk or caused a vehicle to swerve rather than brake (Burrell 2016).

This opacity, sometimes called the "black box problem," created serious evidentiary challenges: it became difficult to diagnose errors, detect bias, or prove that a specific defect in the system was the proximate cause of a harm. In liability contexts, injured parties faced an uphill battle to show "defect" and "causation" when the relevant technical evidence was locked inside an inscrutable model controlled by the producer (Hubbard 2023).

Second, the uncertainty surrounding AI liability disproportionately burdened smaller innovators. Without clear safe harbors or mature insurance markets, the threat of ruinous litigation—whether for copyright infringement or algorithmic discrimination—fell heaviest on those who lacked the massive legal war chests and diversified revenue streams of the tech giants. This dynamic threatened to create a two-tier AI economy, where only entities with sufficient resources to absorb liability risks could fully leverage AI capabilities, leaving smaller organizations and individuals unable to compete. Unless new laws and regulations addressed this gap, the playing field would remain structurally tilted toward incumbents.

Third, AI's general-purpose character meant that governance choices propagated across sectors simultaneously. As AI systems transitioned from narrow applications to general-purpose infrastructure—embedded across finance, logistics, education, healthcare, and national security—regulatory instability threatened not merely a single industry but the trajectory of innovation itself. Issues around data ownership, informed consent, anonymization techniques, data security protocols, misinformation, and transparency all rose to the surface at once. As AI became integral to everything from creative content generation to critical decision-making, the lines between original data, algorithmic processing, and the resulting outputs became increasingly blurred, creating legal challenges that traditional frameworks were not designed to address (Buiten 2019; Schaake 2021).

These challenges interacted with shifts in the structure of the economy itself. During the early 2020s, many of the tightly integrated global supply chains that had blossomed during the earlier era of open markets began to unravel (Alicke et al. 2023). In their place, vertical reintegration emerged across major industries, as firms moved to internalize supply-chain risk and control critical inputs—from batteries and chips to logistics and distribution. Chapter 9 told this story for the AI stack itself.

This wave of vertical integration was driven by a convergence of technological and economic changes. Rapid advances in automation and AI began to erode the core economics of offshoring, especially for routine, low-skill assembly operations that had traditionally been sent overseas. As the unit labor cost benefit of offshoring shrank, domestic production became more viable, and U.S. firms announced a record number of reshoring manufacturing jobs in 2022, citing automation as a key driver (Reshoring Initiative 2023). New technologies like additive manufacturing enabled a shift toward on-demand, on-site production, with investments in the sector accelerating rapidly. Digital logistics—using real-time tracking, digital-twin modeling, and automated warehousing—made geographically closer supply chains more cost-effective by reducing the need for large, precautionary inventories (Manenti 2021).

This technological shift extended beyond manufacturing to services. By 2023, AI-powered virtual assistants were saving global enterprises billions of dollars in customer service costs (Gatford 2025). Major outsourcing firms began embedding AI into their platforms, reducing the need for remote human agents for standard inquiries. The automation of services had a direct impact on offshoring patterns, further reinforcing the trend toward consolidation and internalization.

These shifts matter for AI governance because fragmentation is easier for large firms to absorb. Firms that can internalize compliance, legal risk, and multi-jurisdiction deployment gain an advantage, while smaller innovators face higher effective barriers to entry. The result is that liability, privacy, and security rules often end up determining which firms can deploy AI in high-stakes sectors—healthcare, finance, transportation—rather than simply determining whether AI can be deployed at all.

By 2025, governments seriously grappled with how traditional legal frameworks could address harm caused by AI products and services across applications as diverse as customer service, healthcare, legal services, and autonomous vehicles. While several jurisdictions implemented AI governance mechanisms, others took the drastic step of outright restricting or banning generative AI tools due to data-privacy and security concerns (Naidu et al. 2023). The cross-border flow of data essential to many AI applications raised additional questions about whether international cooperation or harmonization of data privacy regulations might be required (Aaronson and Leblond 2018). Yet the trajectory of the early 2020s pointed away from harmonization and toward fragmentation, a pattern the following sections examine in detail.

The United States: From Creative Destruction to Patrimonial Hybrid

The US entered the AI era with a regulatory inheritance shaped by decades of carefully evaluated innovation policy centered on cost benefit analysis. The Creative Destruction Paradigm combined robust IP protections and openness to free trade, technology standardization efforts sometimes husbanded by the federal government, permissive antitrust, and the provision of critical infrastructure like high-speed wireless internet. This configuration facilitated the rise of the digital platforms that dominated the Third Industrial Revolution. But by 2025, the paradigm had mutated into something less stable and more contingent on political relationships.

The state's role expanded in surprising ways. Industrial policy returned with force. Executive power consolidated at the expense of independent agencies and career civil servants. Antitrust enforcement oscillated between populist aggression and mercantilist forbearance. The resulting hybrid—part market, part patrimonialism—created an environment where outcomes increasingly depended on proximity to political power rather than neutral application of rules.

Industrial Policy

The CHIPS and Science Act, signed in 2022, aimed to strengthen American semiconductor independence by reducing reliance on both Taiwan and China (US Congress 2022). The legislation supported domestic production through subsidies for chip fabrication and tax credits for companies establishing US manufacturing facilities using American-made equipment. It allocated $39 billion in semiconductor subsidies, including $7.9 billion for Intel, and invested significantly in R&D to advance semiconductor technology.

The second Trump administration's posture toward this inheritance was revealing. Though President Trump publicly attacked the CHIPS Act and advocated instead for 100 percent tariffs on foreign-produced chips to promote domestic semiconductor manufacturing, his administration did not abandon the act's core subsidies. Instead, it rebranded the effort. To reduce regulatory burdens and court both domestic and foreign capital, it housed a so-called Investment Accelerator within the Commerce Department, known as the USIA, to oversee and enhance the existing framework. The executive order bringing this project into existence emphasized securing "better deals" and declared that regulatory relief—not just cash subsidies—must accompany these types of initiatives (Trump 2025a).

What distinguished the second Trump administration's approach was not the repudiation of industrial policy but its personalized, transactional character. Industry leaders began to bypass

traditional lobbying in favor of direct relationships with Trump, often making financial commitments or symbolic gestures to win his favor. By early 2025, Trump claimed personal oversight of approximately $1.5 trillion in foreign investment pledges, including TSMC's promise to invest $100 billion (Ip 2025). Domestic firms followed suit: for example, after its CEO Tim Cook presented Trump with a golden plaque, Apple promised to invest an additional $100 billion, on top of investment pledges totaling $400 billion (Bendavid and Bhattarai 2025).

This transactional mode extended to the fine details of export policy. The federal government required Nvidia and AMD to remit 15 percent of their China AI chip sales to the US Treasury as a condition for export licenses—effectively making the government a business partner in their operations. Washington secured "golden shares" and board representation in strategic transactions, including Nippon Steel's acquisition of U.S. Steel and a 15 percent equity stake in MP Materials to control rare earth supply chains (Ip 2025). President Trump personally determined and adjusted tariff levels on imports from nearly every country worldwide, and his administration's treatment of Nvidia's next-generation chip for the Chinese market—the B30A—shifted from hinting at approval to blocking its export after samples had already been sent to Chinese customers (Liu 2025).

The Consolidation of Executive Power

Patrimonialism was not confined to industrial policy. It extended to the basic architecture of regulatory governance. During its first, hyper disruptive year, the second Trump administration pursued an unprecedented drive to bring every corner of the executive branch under unified control. In a January 20th executive order reinstating "Schedule F," it reclassified many long-serving employees as "policy/career" staff subject to removal if they failed to "faithfully implement" White House priorities (Balser et al. 2025). Thousands of provisional hires were

summarily fired, permanent staff were lured out with buyouts, and several inspectors general—Congress's own embedded watchdogs—were dismissed without the statutorily required thirty-day notice. These moves bypassed traditional merit-system safeguards and eroded the neutral competence of the bureaucracy.

The White House also effectively nullified the DoJ's Office of Legal Counsel as an internal check. Historically, OLC opinions have been treated as controlling across the executive branch (see Barron 2010), subject only to review by the Attorney General or President. The administration sidelined OLC memoranda in favor of ad hoc White House legal memoranda, operating under the slogan, "If the President wants it, it's lawful."

Longstanding Supreme Court precedents safeguarding independent-agency tenure were ignored. Although Humphrey's Executor (1935) had insulated multi-member commissions like the National Labor Relations Board and the Merit Systems Protection Board from at-will removal, the second Trump administration fired Democratic appointees to both bodies. When the Supreme Court issued an emergency stay permitting the dismissals to take effect, it sent an unsigned signal that Humphrey's protections were exceptions to, rather than the rule of, Article II removal power (Seila Law 2020).

Beyond personnel and legal theory, the administration wielded its enforcement powers as blunt instruments. The Pipeline and Hazardous Materials Safety Administration initiated only five safety-violation cases, a 92 percent drop from the same period in Trump's first term—after a new legal directive required all enforcement actions to be routed through a single official (Cantwell 2025). EPA enforcement declined significantly, with watchdogs reporting a sharp drop in new civil and criminal cases (Na 2025). At the Office of Federal Contract Compliance Programs inside the DoL, staff were ordered to "cease and desist" enforcement despite binding

civil-rights mandates (Mitchell and Trotta 2025). The DoE was instructed to halt implementation of longstanding water- and energy-conservation standards for household appliances (NAHB 2025).

Antitrust and the Logic of National Champions

The tension between populist instinct and patrimonial logic was nowhere more visible than in antitrust enforcement. The neo-Brandeisian movement, which reached its apogee during the Biden administration, had sought to replace the traditional consumer welfare standard with a broader mandate focused on protecting competition itself and emphasizing the dangers of "bigness" in both economic and political terms (see Khan 2019). This approach garnered bipartisan support: the House Judiciary Committee's 2020 report on Big Tech called for a sweeping overhaul of antitrust law (HJC 2020), and politicians from Democratic Senator Elizabeth Warren to Republican Senator Josh Hawley criticized the economic and cultural power these companies wielded.

The notorious Google search case illustrated the resulting incoherence. In late 2024, a court found Google liable for unlawfully maintaining its dominance in search through exclusionary agreements, most notably its multi-billion-dollar deal with Apple to be the default search engine on the iPhone. In September 2025, the court issued its remedies decision, forbidding Google from making agreements that force partners to give Google the same or better terms than any other search provider gets, and restricting deals that make Google the default search engine on devices (Mehta 2025).

Yet President Trump publicly questioned the wisdom of a structural breakup. Destroying such a "powerhouse," he argued, would only benefit China. "China is afraid of Google," he noted, suggesting that the firm's scale was a strategic national asset that should be disciplined for

"fairness" rather than dismantled (Weiss and Tangalakis-Lippert 2024). This marked a shift from consumer welfare to mercantilist welfare, where monopolies are tolerated if they serve national—or political—interests.

What Remained of the Creative Destruction Paradigm

The consolidation of power witnessed during the first year of the second Trump administration reflected the practical application of a strong unitary-executive orientation—an approach that treats executive-branch administration as subject to direct presidential control rather than insulated, rule-bound expertise. This created friction with the CDP, which depends on stable procedures, professionalized evidence-gathering, and credible commitments that policy will be applied predictably rather than opportunistically. By elevating political loyalty over merit-based safeguards, the administration shifted governance away from neutral competence and toward discretionary, leader-centered implementation. If this trajectory continues, the core test will be whether a complex economy can sustain innovation under institutions that are perceived as contingent—responsive to presidential preference rather than to durable legal and evidentiary constraints.

Elements of the CDP paradigm may yet persist though.

Over its history, litigation proved the primary mechanism for establishing accountability, and that may continue to hold during the AI era. Take the Mobley case, where a job applicant alleged that Workday's AI-driven hiring tool systematically rejected his applications based on race, age, and disability (see Mobley 2025). The court held that Mobley had plausibly alleged that Workday, as an AI vendor, could be treated as an "employer" under civil rights statutes via agency theory, because its tools allegedly performed traditional hiring functions such as screening and rejecting applicants (Tyman 2024). The implication was significant: even as

comprehensive AI legislation stalled at the federal level, courts could constrain AI applications by applying existing civil-rights statutes to vendors whose tools performed core decision-making functions.

And despite an Executive Order preempting state-level AI regulation (Trump 2025b), states nonetheless continued to fill the vacuum left by federal inaction. By late 2025, 47 states had enacted laws specifically targeting deepfakes, focusing on election interference and nonconsensual pornography (Bonatesta 2025). Beyond these commonalities, local political priorities drove diverse mandates. Texas enacted the Responsible AI Governance Act, explicitly banning government agencies from using AI for social scoring (Nahra et al. 2025). Utah's AI Policy Act mandated clear disclosure when a human is interacting with an AI chatbot (Levi et al. 2024). Tennessee's ELVIS Act created the nation's first protections against unauthorized AI cloning of a musician's voice (Damle et al. 2024). California's SB 53 established whistleblower protections for employees at frontier AI labs (Tene et al. 2025).

The result was a fragmented regulatory landscape that large firms were better positioned to navigate than smaller competitors. Companies with resources could maintain compliance infrastructure across jurisdictions, absorb litigation risk, and cultivate political relationships. Those without such resources faced an environment defined more by uncertainty than by durable rules.

China – State-Centric Managerialism

China's approach to AI governance inverted the American model. Where the United States relied on reactive litigation and fragmented state-level rules, Beijing built a proactive, preventative system that required alignment with state goals as a condition of market entry. Where American firms navigated uncertainty by internalizing compliance and cultivating political relationships,

Chinese firms operated within explicit constraints that reduced uncertainty while narrowing the scope of permissible innovation. This configuration—State-Centric Managerialism—reflected not ideological preference alone but the institutional endowments of a political system capable of centralized coordination across economic and security domains.

The Chinese state had long treated technology as an instrument of national power. Since coming to power, President Xi Jinping made "self-reliance" in key technologies a national imperative, arguing that China must "not be forced to beg others for technology" if it is to safeguard its economic and national security (Spegele 2025). Beijing's stated overarching goal was to close the frontier-technology gap with the United States across areas such as semiconductors, artificial intelligence, robotics, and biotechnology. To achieve this, China marshaled an array of state-directed, application-focused industrial-policy tools that blurred the line between public investment and private enterprise.

The Architecture of State-Led Innovation

Beijing's approach was anchored in two flagship plans: the 2017 New Generation Artificial Intelligence Development Plan (AIDP) and the broader "Made in China 2025" roadmap (Webster et al. 2017). AIDP laid out milestones for world-class AI research, domestic chip design, and mass deployment across governance, industry, and defense. "Made in China 2025" sought to replace foreign technology in ten strategic sectors—including semiconductors, robotics, electric vehicles, and biomedicine—boosting the self-sufficiency rate year by year. These plans obligated Beijing to back national champions through subsidies, grants, and preferential procurement, driving an estimated $85 billion investment wave in AI firms and chip startups (Castro and McLaughlin 2024).

The scale of resources mobilized was extraordinary. Chinese R&D spending reached about $500 billion in 2024—roughly triple its 2012 level—and rivaled U.S. outlays on a purchasing-power-parity basis (OECD Statistics 2025). China launched massive investment vehicles to finance rising technologies, most notably the National Integrated Circuit Fund (the "Big Fund"), established to invest in semiconductor and AI hardware firms (Yahoo Finance 2025). Provincial governments and state banks channeled loans and equity into AI startups that aligned with national priorities (Beraja et al. 2025). From 2000 to 2023, China's government-affiliated venture funds poured nearly $200 billion into some 9,600 AI startups, while local investment arms backed challengers like Zhipu AI alongside private giants such as Alibaba and Tencent (Castro and McLaughlin 2024).

These direct investments were enhanced by generous fiscal measures. The centerpiece was a super-deduction for R&D spending, in place since 2007, allowing enterprises to deduct more than 100 percent of their qualifying outlays from taxable income. Initially set at 150 percent, this volume-based R&D tax allowance was raised to 175 percent in 2018 and, as of 2023, stood at 200 percent for most companies (Dezan Shira and Associates 2024). Additional incentives included accelerated depreciation for R&D assets, preferential corporate income tax rates for "High-and-New-Technology Enterprises," and VAT and customs relief on imported R&D machinery. Collectively, these incentives made China's R&D tax subsidy among the world's largest.

Capital flowed toward "national champions" like Huawei, CRRC, and Sinopec, which received $248 billion annually in industrial subsidies. Between 2020 and 2024, state-controlled banks lent an additional $1.9 trillion to industrial borrowers (Casanova and Miroux 2018). Beijing also turned to massive government procurement contracts to endow Chinese firms with

advantages in reaching economies of scale and used the Belt and Road Initiative to secure access to energy-related natural resources while weaponizing supply chains through control of critical minerals.

DeepSeek and the "Swarms Beat the Titan" Strategy

The fruits of this investment became visible in early 2025, when DeepSeek's release of competitive AI models at a fraction of Western training costs "jolted American competition into overdrive" and became the focal point of Beijing's national AI plan. Premier Li Qiang reportedly told officials, "China finally has a model it can be proud of," and President Xi Jinping personally convened a meeting with DeepSeek's founder to mobilize the industry (Chin and Huang 2025).[285]

DeepSeek's open-sourced models were rapidly integrated into the IT infrastructure and workflows of local governments and state-owned companies across China. This diffusion strategy served both economic and governance objectives: it reduced dependence on foreign

[285] Beijing designated DeepSeek as a "national treasure." The Chinese government implemented travel restrictions on some employees involved in AI model research and development, requiring them to hand in their passports due to concerns about confidential information, potentially including trade and state secrets, being leaked. The local government of Zhejiang Province, where DeepSeek's parent company is headquartered, began screening potential investors before allowing them to meet with company leaders, registering investment inquiries with the provincial Communist Party committee. Headhunters were contacted by government officials and asked not to poach talent from DeepSeek (Chin and Huang 2025).

models while ensuring that the AI systems embedded in Chinese institutions operated within parameters acceptable to the state.

Finally, to compensate for its lack of cutting-edge chips, China pursued a "swarms beat the titan" strategy (Chin and Huang 2025). This involved building AI supercomputers that bundled together thousands of less powerful, domestically produced accelerators—like Huawei's Ascend chips—to match the performance of a smaller number of Nvidia's "titan" GPUs. Beijing strongarmed local governments to shower data centers with deep electricity subsidies if they agreed to use these domestic chips to address the fact that this "swarm" approach consumed significantly more power.

Regulatory Integration: AI as Political Instrument

China's AI-specific regulations reflected the same logic of alignment and control. Leading into 2025, China had rolled out some of the world's earliest AI-specific regulations. The *Provisions on the Administration of Deep Synthesis Internet Information Services*, which went into effect in 2023, regulated "deep synthesis" services—AI-generated or manipulated audio-visual content—by requiring providers to label synthetic media and prevent its use to spread illegal or misleading information (CAC 2022). Also in 2023, the Cyberspace Administration of China and six other ministries issued the *Interim Measures for the Management of Generative Artificial Intelligence Services* (IMGAIS), which governed public-facing generative-AI by requiring security assessments, algorithm filings, content controls, and multi-agency supervision for systems offered to users in China (CAC 2023).

The IMGAIS framework codified stringent content requirements. Providers must ensure that model outputs adhere to China's censorship regime and "core socialist values," filtering or preventing content that is politically sensitive, violent, sexual, false, or otherwise illegal under

Chinese law. If a model generates prohibited material, the service provider—not the user—is held responsible and must correct the problem (ibid). This diverged sharply from the U.S. intermediary-immunity impulse often associated with "Section 230": Chinese generative-AI platforms were treated more like accountable publishers than neutral conduits.

The rules also introduced security assessments and algorithm-filing obligations. Providers must disclose basic information about model architecture, training-data categories, and intended use cases, giving authorities leverage to intervene if outputs are deemed unsafe or politically destabilizing (White & Case LLP 2025). Major platforms such as Baidu and Alibaba were only allowed to deploy ChatGPT-style services after demonstrating robust censorship of "sensitive topics" and safety filters.

Transparency and labeling duties were strengthened and then generalized. Building on the 2022–23 deep-synthesis rules, which mandated labels for AI-generated audio-visual material, China adopted dedicated *Measures for the Administration of Labels for Artificial Intelligence-Generated Content* in March 2025, requiring AI-generated text, images, video, and audio that could be mistaken for real content to carry clear labels (Cyberspace Administration of China 2025). The accompanying national standard specified a dual system of explicit indicators—such as on-screen marks, watermarks, or audio tones—and implicit technical markers embedded in metadata or digital signatures (White & Case LLP 2025). In diffusion terms, these labeling rules turn governance into infrastructure: they make compliance legible and enforceable at the distribution layer.

Chinese courts reinforced these regulatory expectations through emerging case law. In February 2024, the Guangzhou Internet Court issued the so-called Ultraman decision, widely described as the first effective judgment on generative-AI copyright infringement. The court held

an AI-image platform liable because its system generated pictures of the famous Ultraman character that were substantially like the copyright holder's works. It emphasized that providers have a "duty of care" to implement keyword filtering, monitoring, and complaint mechanisms—effectively articulating a negligence standard for AI services (Song and Wang 2024). The court's reasoning underscored that, unlike in the U.S., Chinese providers cannot invoke broad immunity when their models output infringing content.

Winning the Open-Source Race

Perhaps the most consequential dimension of China's AI strategy was its success in open-weight model development. In 2025, Chinese tech firms began to win a key part of the AI race: the battle for open-weight dominance. While American companies were widely seen as offering the most powerful proprietary models, Chinese firms increasingly outcompeted their U.S. rivals in "open-weight" AI—technology that anyone can freely download and build upon. This developer-permissive approach, which makes a model's learned parameters available for modification, was championed by Chinese firms with high-performing models like DeepSeek R1 and Alibaba's Qwen3 (Huang 2025).

Chinese firms secured a significant reversal from the previous year, when U.S.-based Meta offered the world's best open model. By October 2025, the top-ranked open-weight models on LMArena, a site that ranks models based on blind user preference tests, were dominated by Chinese companies like DeepSeek, Z.ai, and Alibaba (Fu 2025). This popularity was confirmed on Hugging Face, the leading platform for AI developers: in early November of 2025, DeepSeek-R1's model card showed about 12.8k likes—more than twice Meta's most-liked recent Llama-3.x instruct release, which garnered about 4.9k likes (Hugging Face 2025b; Hugging Face 2025c).

Because these Chinese models excelled at local-language tasks and could be run in-house, they rapidly became a default choice for enterprises prioritizing data control. This adoption included multinational banks like HSBC and Standard Chartered, industrial giants like Saudi Aramco, and even major U.S. cloud vendors like AWS, Microsoft, and Google (Lin et al. 2025). The trend directly challenged the Trump administration's AI strategy, leaving the U.S. scrambling to respond with new initiatives like the American Truly Open Models (ATOM) Project (ATOM Project Substack 2025).

This competition for open-weight dominance was being fought most intensely in the lucrative market for AI coding assistants. In 2025, Chinese startups like MiniMax, Zhipu, and Moonshot AI released a wave of new models that were not only high performing but radically cheaper than U.S. alternatives (Osawa 2025). MiniMax's M2 model ranked highly on developer leaderboards and was priced at just 8 percent of the cost of Anthropic's competing Claude Sonnet model. This aggressive strategy was driven by the need to find paying customers outside of China, where fierce domestic price wars made it difficult to generate revenue. The approach was proving successful: by late 2025, U.S. cloud company Vercel had partnered with Zhipu, and celebrated venture capitalist Chamath Palihapitiya had migrated workloads to Moonshot AI's Kimi K2.

The Tradeoffs of State-Centric Managerialism

China's approach produced a distinctive configuration. By integrating AI governance into centralized political oversight, Chinese authorities reduced uncertainty about permissible uses while narrowing the scope of acceptable experimentation. Algorithmic filing requirements, content controls, and security assessments imposed uniform expectations across firms, but they

also aligned innovation trajectories with state priorities. This model facilitated rapid diffusion in targeted applications while constraining development in politically sensitive domains.

The consequence was not uniform stagnation or acceleration, but a channeling of innovation along paths compatible with existing political objectives. State-led projects achieved impressive results: by 2025, China accounted for about 54% of global industrial robot deployments, and had surpassed the U.S., Germany, and Japan in robot density (Casado and Neuberger 2025). Chinese firms like Unitree produced advanced humanoid robots for as little as $5,900, demonstrating they were not cheap knockoffs but sophisticated machines (Robotics Tomorrow 2025). This created a strategic risk for the United States, which lacked a comparable robotics ecosystem.

Yet China's approach also carried potential long-term costs. By requiring ideological alignment as a condition of market participation, China foreclosed certain innovation trajectories that might have emerged from less constrained experimentation. The emphasis on predictability came at the expense of the open-ended exploration that had characterized earlier phases of digital innovation elsewhere. Whether this tradeoff would prove advantageous or limiting depended on developments that remained uncertain, including whether the technologies most consequential for economic growth would emerge from state-directed channels or from the distributed experimentation that centralized systems struggle to accommodate.

Europe – The Regulatory-First Gamble

The European Union pursued a third path, distinct from both the American hybrid and Chinese state-centrism. Where the United States relied on ex post litigation and fragmented state action, and China required ex ante alignment with political objectives, Europe sought to establish comprehensive rules before AI systems reached the market. This Regulatory-First model

reflected the EU's institutional strengths—high capacity for rulemaking, normative coherence around rights and safety, and experience projecting standards globally through market access. But it also exposed the EU's weaknesses: fragmented capital markets, sluggish commercialization, and deepening dependence on foreign AI infrastructure. The gamble was that regulatory clarity and market access would pull innovation and infrastructure into Europe faster than Europe's structural constraints would push them away.

By 2025, the gap between Europe's regulatory ambition and its industrial capacity had become a source of acute political tension. The continent faced serious structural headwinds. An aging population was shrinking the future workforce. Welfare states were under strain. Tax burdens were higher than in competitor economies. Red tape had doubled in volume since 2010. Energy prices—crucial for compute-heavy AI—were multiples of US industrial power costs, with German rates roughly three times American levels, undercutting the siting of data centers and advanced manufacturing (Luhnow and Fairless 2025).

The scale of Europe's innovation deficit was stark. Over the last half-century up to 2025, the United States produced 241 new $10-billion-plus public companies from scratch, versus only 14 in Europe, while EU labor productivity per hour slipped from roughly 95 percent of U.S. levels in the late 1990s to under 80 percent by 2025 (Fairless and Luhnow 2025). As of early 2025, there were about 100 European "unicorns"—roughly one-third of a trillion dollars in combined value—far behind the U.S. and China in both counts and valuations (CB Insights 2025d; Fairless and Luhnow 2025). By 2025, only four of the world's top 50 publicly traded tech companies were European, despite the EU representing about 14 percent of world GDP on a purchasing-power-parity basis (Draghi 2024). The problem was not rulemaking capacity; it was the difficulty of translating scientific capability into scaled firms and scaled infrastructure.

The semiconductor sector illustrated the problem with particular clarity. While the region still hosted critical champions like ASML and ARM, Europe's share of global manufacturing capacity had eroded (Saint-Martin 2024). The EU itself set a goal to double its global chip share to 20 percent by 2030—an implicit acknowledgment that it hovered near about 10 percent—and Europe had no prominent pure-play foundry competing with TSMC or Samsung at advanced nodes (European Court of Auditors 2025). Investment in domestic chipmaking lagged the scale of rival programs, with the U.S. CHIPS and Science Act appropriating $52.7 billion while estimates of China's support exceeded $150 billion over 2014–2030 (Blevins et al. 2023). In a fragmented world of export controls and competing blocs, this meant Europe's governance model increasingly rested on infrastructure and supply chains it did not fully control.

The AI Act: Risk Classification and the Brussels Effect

Against this backdrop, the EU Artificial Intelligence Act represented an attempt to convert regulatory capacity into competitive advantage. Finalized after trilogue negotiations concluded in December 2023, the Act established a comprehensive "risk-based" regime (European Parliament 2024). It banned certain unacceptable uses of AI, placed strict pre-approval controls on "high-risk" systems used in areas like healthcare or critical infrastructure, and imposed transparency duties on lower-risk applications like chatbots (European Union 2024).

The Act categorized AI applications into four tiers: unacceptable risk, which encompassed banned applications like social scoring systems; high risk, covering systems affecting health, safety, or fundamental rights; limited risk, requiring transparency; and minimal risk, with few or no requirements (European Parliament 2024). Key provisions included targeted bans (including limits on real-time biometric identification in public spaces) and transparency

and disclosure rules for chatbots, generative AI, and deepfakes. Non-compliance could result in fines up to €35 million, or 7 percent of global annual turnover, whichever was higher.

A new European AI Office was created within the European Commission to enforce these rules, particularly for powerful general-purpose AI models. To guide companies during the transition, a voluntary Code of Practice was developed with over a thousand stakeholders, offering a path for developers to demonstrate compliance with key requirements like model documentation, security controls, and risk evaluation before formal standards were finalized (European Commission 2025a).

The strategic logic underlying this framework was the "Brussels Effect"—the tendency for EU regulations to become de facto global standards as multinational firms adapt their products and processes to retain access to the European market. Because many non-EU suppliers adapted to these harmonized protocols, the framework often projected outward, shaping practices well beyond Europe and encouraging firms to adopt the strictest common standard as a default to avoid maintaining region-specific variants (Engler 2023). In anticipation of the AI Act, American and Chinese firms expanded AI governance, conducted risk assessments, trained staff on data-handling and model-use policies, and staged deployments until they were confident of alignment with European standards before scaling (ibid).

Workday, a US-based enterprise software company headquartered in Pleasanton, California, provided a concrete example. As a provider of AI-enhanced HR and finance platforms used by global clients including many in the EU, Workday proactively adapted to EU standards to maintain market access. The company expanded its AI governance framework by integrating ISO/IEC 42001 for AI management systems, which emphasizes structured risk management and documentation. This involved creating detailed model cards for transparency,

embedding bias-mitigation protocols in development pipelines, and adding watermarking for generated content to align with transparency expectations under limited-risk rules (see Workday 2023).

The Competitiveness Tension

Yet even as the Act moved toward finalization, a deep rift emerged between the EU "regulators" in Brussels and the "industrialists" in key member states. Leaders like French President Emmanuel Macron and German Chancellor Olaf Scholz actively pushed back against the Act's strictest provisions for "general-purpose" AI models (Henshall 2023). Their fear was that excessive red tape would strangle Europe's own nascent champions—such as France's Mistral AI and Germany's Aleph Alpha—before they could compete with U.S. giants (Davies 2023).

This political tug-of-war mirrored the American tension between safety and innovation and, by 2025, highlighted a growing consensus across the Atlantic that strategic competitiveness might require tempering the impulse to regulate (WEF 2025). The evidence for European disadvantage was difficult to ignore. Germany's Aleph Alpha, for example, pivoted away from the frontier-model race toward enterprise solutions and consulting-style deployments, while many promising teams sought U.S. capital markets, looser regulation, and stock-option incentives more conducive to rapid scaling (Lomas 2024; Fairless and Luhnow 2025).

Capital markets presented a fundamental weakness. In 2025, European venture-tech investment remained roughly one-fifth the U.S. level, as much financing still originated from banks or large pension funds, leaving a "missing middle" of dynamic, growth-oriented capital (Fairless and Luhnow 2025). Historically, European capital markets—especially in Germany—provided less venture capital compared to the U.S. and even China, partly due to more conservative banking structures, fewer tax incentives for entrepreneurship, and differences in

equity culture (Veugelers 2018). While the U.S. fostered high-growth startups across the Third and Fourth Industrial Revolutions, including AI powerhouses such as Anthropic and OpenAI, European innovators struggled with early-stage funding (Schneider 2024).

Regulatory and cultural barriers compounded these challenges. Europe's complex multi-jurisdictional market and stricter labor frameworks made cross-border scaling and hiring more difficult, while historically less-friendly equity rules dampened stock-option use (Baroudy et al. 2020). In several countries, stock options were even taxed at grant rather than at exercise or sale—further reducing their attractiveness (Palmer 2019). Companies reported spending about 40 percent of IT budgets on compliance with EU regulations, and roughly two-thirds of firms said they still struggled to understand their obligations under the EU AI Act (Duffer et al. 2025).

Stepping back, the deeper difference is how each bloc turns governance into implementation. A useful way to see these governance models is through their approach to standardization. The EU increasingly tries to convert regulation into a single technical rulebook, using harmonized standards that make compliance—and thus market access—legible and enforceable at scale. China pushes in the opposite direction, treating compliance as a precondition for market entry and building binding obligations into a vertically integrated regulatory stack. The United States has relied more heavily on ex post enforcement and voluntary frameworks, which preserves experimentation but can leave diffusion hostage to liability uncertainty and patchwork adoption. In diffusion terms, these are not just legal differences: they shape the transaction costs of adoption and the extent to which AI tools can travel across borders without being rebuilt.

The Sovereign Compute Response

Recognizing that legal safe harbors alone were insufficient, the EU moved to back its rules with hard power—major investment commitments aimed at achieving "sovereign compute," a domestic, EU-controlled infrastructure free from reliance on foreign platforms. This strategic pivot reflected growing anxieties over Europe's deep reliance on American cloud providers (Johnson and Hay 2025). By 2025, there were clear signs that European companies and governments were souring on their dependence on U.S. "hyperscalers" like Amazon Web Services, Microsoft Azure, and Google Cloud. Denmark's two largest municipalities, for example, moved to end use of Microsoft systems amid sovereignty concerns (Desmarais 2025). These anxieties were amplified by the return of the Trump administration, renewed uncertainty over transatlantic data-sharing arrangements, and fears tied to US surveillance authorities such as the CLOUD Act and FISA §702 (Johnson and Hay 2025).

The push for sovereignty was intellectually framed by "EuroStack," a programmatic blueprint commissioned by European institutions (Bria et al. 2025). The EuroStack report outlined a strategy for sovereign digital infrastructure anchored in interoperability, sustainability, and treating "data as a common good," calling for tools like a "Buy European" act, close to €300 billion in sovereign tech investment, and a "sovereign AI cloud" marrying public supercomputers with decentralized systems.

By 2025, these aspirations were mirrored in concrete EU programs. The EuroHPC Joint Undertaking (JU), an agency that began upgrading its traditional scientific supercomputers with thousands of GPUs, sought to create "AI factories" capable of supporting European model development (EuroHPC JU 2024). This effort was complemented by the European Commission's "InvestAI" plan, a €200 billion initiative designed to fund these "AI

gigafactories" and provide European startups with the domestic capacity needed to build large-scale models (Kyosovska and Renda 2025).

By mid-2025, scattered initiatives were consolidated into a unified industrial strategy known as the "AI Continent Action Plan." Launched in April, this roadmap explicitly aimed to transform Europe into a global AI leader by integrating infrastructure investment with regulatory certainty and tying AI compute "gigafactories" and AI Factories to a new InvestAI financing facility (European Commission 2025a). It was promptly followed by the "Apply AI Strategy," a targeted push to fast-track AI adoption across Europe's legacy industrial base—from automotive manufacturing to healthcare—addressing the continent's persistent productivity gap (ibid).

Antitrust as Industrial Policy

Europe also wielded competition policy as an instrument of industrial strategy, though with ambiguous results for domestic innovation. Historically, European antitrust authorities—though roughly guided by a consumer welfare standard, like the United States—were more vigilant about preventing market dominance and less inclined to promote innovation for innovation's sake (Schneider 2024). In practice, this emphasis on curbing concentration tended to discourage "winner-takes-most" platform dynamics. For AI, that posture carried a tradeoff: smaller user bases and more fragmented data pools limited the scale advantages that frontier model training increasingly rewards (Veugelers 2018; Chapter 7 of this book).

Over the 2020s, the EU supplemented traditional, case-by-case enforcement with ex ante rules for designated "gatekeeper" platforms under the Digital Markets Act (DMA). [286] In 2024,

[286] Under the DMA, these core obligations are designed to ensure contestability and fairness: "steering" prohibits gatekeepers from preventing app developers and business users from

the European Commission opened its first non-compliance investigations against Alphabet, Apple, and Meta; by 2025, it formally found Apple (anti-steering) and Meta (pay-or-consent) in breach, and pressed Google to revise Play Store terms on fees and external offers (European Commission 2025). The DMA also sat atop a decade of headline abuse decisions—including the Android case—where EU courts largely upheld multibillion-euro fines (EU General Court 2022), signaling that Brussels was serious about disciplining dominant digital platforms.

A series of decisive enforcement actions against Big Tech firms across 2024 and 2025 demonstrated the EU's ability to discipline global platform firms—and, in doing so, reinforced the broader logic of the Brussels Effect. The European Court of Justice upheld a €2.4 billion fine against Google for self-preferencing its Shopping service and ordered Apple to pay €13 billion in back taxes to Ireland, overturning a lower court victory for the company (CJEU 2024a, 2024b). The EU General Court annulled a separate €1.5 billion fine against Google's AdSense business (EU General Court 2024). The European Commission fined Apple €500 million for anti-steering violations in its App Store and Meta €200 million for its "pay or consent" data model (European Commission 2025b). Perhaps the most significant victory came not from a fine but from a structural concession. Under sustained EU pressure, Microsoft agreed to offer Office and

---

informing their customers about alternative, often cheaper, offers outside the gatekeeper's ecosystem; "self-preferencing" forbids gatekeepers from ranking their own products or services more favorably (e.g., in search results) than similar offerings from third-party competitors; and "access" and interoperability mandates require gatekeepers to allow third-party service providers and hardware manufacturers access to the same operating system, hardware, or software features that the gatekeeper uses for its own services (EU 2022).

Microsoft 365 suites without Teams at reduced prices and to improve interoperability and data portability (European Commission 2024).

Europe's Uncertain Future

Yet despite these efforts—Europe's regulatory buildout, the AI Act's implementation push, and the turn toward sovereign compute—Europe's innovation comeback had "barely scratched the surface" relative to the U.S. and China. The fundamental tension remained unresolved: Europe possessed the institutional capacity to regulate but struggled to generate the dynamic ecosystems that produced frontier AI systems. In 2025, none of the top ten global investors in quantum computing were European, and only a handful of firms on the continent pursued frontier-scale AI models (Fairless and Luhnow 2025).

To be frank about Europe's position circa 2025: it was militarily weak, had a stagnant economy with an almost non-existent AI sector, moved sluggishly due to the constant need for compromises baked into the European Union's Byzantine political structure, and possessed a relatively open economy that was extremely vulnerable to tariffs and other trade and geopolitical disruptions (Authers 2026). The continent's slippage showed up beyond spreadsheets: in 2025 the US executed 100-plus orbital launches, China 40-plus—Europe only four, after a year of relying on SpaceX to loft critical infrastructure (Luhnow and Fairless 2025).

The Regulatory-First model thus represented a gamble whose outcome remained uncertain. If global demand for trustworthy, rights-respecting AI systems grew, Europe's head start in compliance infrastructure might prove valuable. If speed and scale determined market outcomes alone, and the AI frontier kept moving at lightning speed, the EU risked becoming an ignorable, albeit pesky, rule-maker—judging AI systems built elsewhere while its own firms struggled to compete at the frontier.

CONVERGENCE TOWARD FRAGMENTATION

An American mutation toward a patrimonial hybrid, the Chinese consolidation of state-centric managerialism, and the European pursuit of a regulatory-first strategy: while these configurations differed in their institutional origins, mechanisms of control, and characteristic tradeoffs, beneath these differences lay a striking convergence. Across the early 2020s, all three blocs moved toward greater statism and technological mercantilism. The result was not harmonization around a single model of AI governance but fragmentation into competing, gated ecosystems.

The drivers of this convergence were structural rather than ideological. AI development depended on inputs—advanced chips, training data, cloud infrastructure, specialized talent—that were unevenly distributed and increasingly subject to strategic competition. Control over these inputs conferred leverage; dependence on foreign suppliers created vulnerability. As awareness of these dynamics spread, policymakers across systems reached similar conclusions: AI was too consequential to be left to markets alone, and reliance on foreign infrastructure posed unacceptable risks. The instruments varied, but the direction of travel was consistent.

This structural logic was reinforced by the collapse of the post-2008 globalization consensus. On April 2, 2025, President Donald Trump issued a proclamation declaring "Liberation Day" and imposing sweeping tariffs under emergency powers (Barath et al. 2025). The administration justified these measures under the International Emergency Economic Powers Act, bypassing normal trade-remedy procedures and the WTO framework entirely (Trump 2025c). Retaliation followed, and multiple analyses estimated that the effective U.S. tariff rate jumped overnight to levels not seen since the Smoot-Hawley era (Ip 2025; Irwin 1996). Taken together, these protectionist measures accelerated the shift away from the postwar rules-based trade system (Barath et al. 2025).

The consequences quickly appeared at the product layer. Firms could no longer assume that systems developed for one market would be deployable elsewhere without significant adaptation, and global companies increasingly faced pressure to maintain region-specific variants—a "GDPR AI" for Europe, a "State-Approved AI" for China, and a "Liability-Shielded AI" for the United States (Engler 2023).

Fragmentation also ran through infrastructure. Cross-border data flows were constrained by sovereignty requirements, and cloud computing became a vector of geopolitical concern as governments questioned whether sensitive workloads should run on foreign-controlled infrastructure (O'Hara and Hall 2025). Even open-source collaboration—long treated as borderless—was drawn into strategic competition, as the battle between American and Chinese models for developer adoption took on national significance (Fu 2025).

The cumulative effect was the enclosure of what had, until then, been a digital commons. For three decades, the internet operated on a rough consensus of borderless exchange: data, capital, and code flowed relatively freely across a single global supply chain. That era was ending. While physical borders constrained semiconductor manufacturing and energy grids, legal borders enforced data sovereignty and allocated liability, and ideological borders governed the purpose and control of machine intelligence.

What, then, are the stakes of this fragmented landscape? And what does it portend for the future of innovation?

Beyond the AGI Race

That implementation race, in turn, depends on a balance between centralized infrastructure ("Titans") and distributed experimentation ("Swarms").[287]

The early AI era demanded massive, centralized capital to build the necessary physical foundations: semiconductor fabrication plants, data centers, cloud infrastructure, and energy systems. These were the "Titans"—the hyperscale platforms and state-backed champions capable of mobilizing resources at a scale no startup or research consortium could match. Without the Titans, there would be no hyperscale training and deployment at the frontier, and no infrastructure capable of serving billions of users.

However, even if America maintains a meaningful lead in raw model performance by restricting chips and other inputs, foreign competitors may close much of the performance gap at far lower cost thanks to the steep returns on modest hardware investments and continual efficiency gains in both algorithms and semiconductor design. DeepSeek's emergence demonstrated that a relatively small team, operating under resource constraints imposed by export controls, could achieve frontier-competitive performance through architectural innovation and efficiency optimization. The explosion of open-weight models in 2025 showed that value creation was not confined to the largest laboratories. Startups, researchers, and independent

---

[287] The "swarms beat the titan" metaphor circulated in contemporary commentary on China's AI strategy during 2025—especially in discussion of system-level approaches that aggregate many weaker chips and optimize software/architecture under export-control constraints (see Chin and Huang 2025). I use the shorthand "Titans and Swarms" here as an organizing device to capture a broader political-economy tension: the virtues of a centralized, capital-intensive infrastructure versus distributed experimentation and application-layer innovation.

developers—the “Swarms”—built applications the “Titans” had not anticipated, discovered use cases centralized planning would not have identified, and pushed the technology in directions no roadmap had specified.

Much public discourse about AI competition focused on the race to achieve artificial general intelligence, a hypothetical system capable of matching or exceeding human cognitive abilities across all domains. Headlines tracked which laboratory had achieved the latest benchmark, which model could pass the most demanding examinations, and which nation was “winning” the race to this technological frontier. The framing captured genuine stakes: AGI, if achieved, would represent a discontinuity in human history, with consequences difficult to foresee. And it was not clear that in this race the Swarms could match the Titans on their own terms.

Yet this focus on the furthest reaches of the technological frontier may have been misplaced. The previous chapter argued that what will ultimately determine economic outcomes is not which nation possesses the most advanced model in a research setting, but which can most effectively integrate AI into routine production processes and internal administration. Consistent with what I’ve argued in this book in general, therefore, the more immediate and consequential contest was what some observers called the “AI implementation race”: the practical challenge of embedding today’s AI into every factory, hospital, school, and military unit to unlock real-world productivity gains (Alhassani and Bak 2025). The winner of the AI race would not necessarily be the nation that first achieved AGI in a laboratory. Victory might instead belong to the economy that successfully diffused AI tools—perhaps more primitive ones—into the physical bedrock of society.

The Fate of the Creative Destruction Paradigm

What, finally, may come of the Creative Destruction Paradigm—the configuration that had governed American innovation policy for four decades, shaped the global digital economy, and indirectly ushered in AI? Was it dead?

The evidence suggested it was mutating rather than disappearing, increasingly distorted by a new patrimonialism, where innovation was encouraged but the state increasingly picked winners in strategic sectors and courts settled the rules of fair play. Markets remained the primary arena of competition, but access to that arena was mediated by political relationships, regulatory discretion, and the capacity to absorb legal risk. The paradigm's core commitments—to decentralized innovation, neutral rules, and ex post correction rather than ex ante permission—persisted in form but were hollowed out in practice.

Ultimately, while the judicial branch may ensure that the CDP breaks in one clear direction, its role as arbiter may take time—and may work in unexpected ways. Take the Supreme Court's Loper Bright decision, which overruled Chevron deference and required courts to independently interpret ambiguous statutes rather than deferring to agency expertise (Loper Bright 2024). This shift cut in multiple directions: it constrained executive overreach in one sense, but limited the flexibility that had allowed agencies to adapt regulations to changing technological conditions in another. The major questions doctrine, which required clear congressional authorization for rules of vast economic or political significance, imposed similar constraints (West Virginia et al. 2022). Whether these doctrines serve as checks on patrimonial drift or merely transfer discretion from agencies to courts remains to be seen.

An Unfinished Revolution

By the end of 2025, the Fourth Industrial Revolution was unfolding in a world that looked increasingly unlike the one that had incubated the Third. The post-Cold War consensus

on open markets, integrated supply chains, and regulatory convergence had fractured. The institutions that governed innovation—from antitrust enforcement to intellectual property protection to administrative rulemaking—were under strain, their legitimacy contested and their capacity uncertain.

Yet the technology continued to advance. Models grew more capable. Applications proliferated. The gap between what AI systems could do in principle and what they were permitted to do in practice widened in some domains and narrowed in others. The revolution was real, but its trajectory was not determined by technology alone. It was shaped by governance choices—choices made under constraint, reflecting the political economies in which they were embedded, producing equilibria that were stable but not permanent.

The fragmented landscape that emerged by 2025 was probably not the end of the story. New technologies could shift the balance between Titans and Swarms. New crises could accelerate or reverse the drift toward mercantilism. New political coalitions could double down on patrimonialism or, alternatively, resuscitate the original version of the Creative Destruction Paradigm. As Schumpeter emphasized, no matter one's best laid plans, the future is never fully foreclosed.

What was clear in January of 2026 is that the choices made in these years, if not months, would cast long shadows. The infrastructure being built, the precedents being set, the coalitions being formed—these would shape the possibilities available to future policymakers and innovators. The Fourth Industrial Revolution would be defined not only by the capabilities of machines but by the political economies that determined how those capabilities were developed, deployed, and governed. The revolution was underway. Its outcome remained to be written.

BIBLIOGRAPHY

3GPP. 2019. "Third Generation Partnership Project: Working Procedures." 3GPP Organizational Partners. https://www.3gpp.org/ftp/information/working_procedures/3gpp_wp.htm

3GPP. 2021b. Technical Specification Group working methods (TR 21.900). Version 17.0.0. Valbonne, France: 3GPP Organizational Partners. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=555

3GPP. 2024. "About 3GPP." https://www.3gpp.org/about-us

A&P 1949. United States v. New York Great A&P Tea Co., Civ. No. 52-139. US District Court, S.D. New York (filed Sept. 15). https://law.justia.com/cases/federal/district-courts/FSupp/67/626/2311543/

AAR. 2025. Staggers Rail Act of 1980. AAR. https://www.aar.org/issue/staggers-act-of-1980/

Aaronson, Susan Ariel, and Patrick Leblond. 2018. "Another Digital Divide: The Rise of Data Realms and Its Implications for the WTO." Journal of International Economic Law 21(2): 245-272.

ABA 1969. Commission to Study the Federal Trade Commission. 1969. Report of the ABA Commission to Study the Federal Trade Commission. https://www.uschamber.com/assets/documents/Primer-on-The-FTC's-New-Section-5-Guidance.pdf

ABA. 1956. "Report and Recommendations Presented to House of Delegates by Section of Antitrust Law at Midwinter Meeting, Chicago, Illinois, February 20–21, 1956." https://www.justice.gov/atr/merger-guidelines/overview

Abbate, Janet. 1999. Inventing the Internet. Cambridge, MA: The MIT Press.

Abelson, Hal 1994. Summary of In re Alappat. Prepared by counsel for Tektronix, Inc. Hosted at MIT CSAIL. https://groups.csail.mit.edu/mac/classes/6.805/assorted-short-pieces/alappat-summary.html

Abernathy, Frederick, John Dunlop, Janice Hammond, and David Weil. 1999. A Stitch in Time. New York: Oxford University Press.

Accredited Standards Committee X9 (ASC X9). 2022. Financial Services Data Security Standards. ASC X9.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In Handbook of Labor Economics, Volume 4. Edited by Orley Ashenfelter and David Card. Handbook of Labor Economics. Amsterdam, Holland: Elsevier, 1043–1171.

Acemoglu, Daron. 2002. Directed Technical Change. Review of Economic Studies 69(4): 781–809.

Acemoglu, Daron. 2009. Introduction to Modern Economic Growth. Princeton: Princeton University Press.

Acemoglu, Daron. 2024. “The Simple Macroeconomics of AI.” MIT Economics Working Paper / Economic Policy. https://economics.mit.edu/sites/default/files/2024-04/The%20Simple%20Macroeconomics%20of%20AI.pdf

ACT. 2021. Digital Markets Act Position Paper. The App Association. https://actonline.org/wp-content/uploads/ACT-The-App-Association-DMA-Position-Paper-March-.pdf

Actian. 2025. Overcoming Manufacturing’s Biggest Data Challenges: From Silos to Trust. White Paper. https://www.actian.com/wp-content/uploads/2025/02/overcoming-manufacturings-biggest-data-challenges-from-silos-to-trust.pdf

Addyston Pipe. 1898. United States v. Addyston Pipe and Steel Co., No. 498. US Court of Appeals, Sixth Circuit, February 8. https://law.justia.com/cases/federal/appellate-courts/F2/85/271

Adeboye, Daniel. 2025. "How much does an NVIDIA H100 GPU cost?" Northflank Blog. https://northflank.com/blog/how-much-does-an-nvidia-h100-gpu-cost

Adelman, Morris Albert 1959. A&P: A Study in Price-Cost Behavior and Public Policy. Cambridge, MA: Harvard University Press.

ADQ. 2025. ADQ and Energy Capital Partners to establish a USD 25 billion US-based investment partnership. ADQ. https://www.adq.ae/newsroom/adq-and-energy-capital-partners-to-establish-a-usd-25-billion-us-based-investment-partnership-focused-on-developing-new-power-generation-to-serve-the-growing-electricity-needs-of-data-centers/

Affinius Capital. 2025. “From Cloud to Capital.” PREA Quarterly Global (Spring 2025), Exhibit 2. Affinius Capital. https://affiniuscapital.com/app/uploads/Spring_2025_PREA_Quarterly_Global.pdf

Aggarwal, Sanjay, and Betsy Mulé. 2025. State of Robotics. F-Prime Capital. https://docs.google.com/presentation/d/1jMNIjt3rU2lQ_WNlKSy2UUQkE_Dr9FJhMecAnjCQnw8/edit?slide=id.p#slide=id.p

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth Through Creative Destruction." Econometrica 60(2): 323-351.

Aghion, Philippe, and Peter Howitt. 1998. "On the Macroeconomic Effects of Major Technological Change." In General Purpose Technologies and Economic Growth, edited by Elhanan Helpman. Cambridge, MA: MIT Press, 121–144.

Aghion, Philippe, Ufuk Akcigit, and Peter Howitt. 2014. What Do We Learn from Schumpeterian Growth Theory? In Handbook of Economic Growth, Vol. 2, edited by Philippe Aghion and Steven N. Durlauf. Amsterdam, Holland: Elsevier, 515–563.

AGNCSAL. 1955. Report of the Attorney General's National Committee to Study the Antitrust Laws. Washington, DC: U.S. Government Printing Office.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2023. Similarities and Differences in the Adoption of General Purpose Technologies. NBER Working Paper No. 30976. Cambridge, MA: National Bureau of Economic Research.

Ahmed, Nur and Muntasir Wahed. 2020. "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research." arXiv preprint arXiv:2010.15581. https://arxiv.org/abs/2010.15581

AInvest. 2025. "Intel's Layoffs: A Necessary Reset or a Last Gasp?" AInvest. https://www.ainvest.com/news/intel-layoffs-reset-gasp-2506/

Akhtar, Shayerah Ilias. 2013. Trade reorganization: Overview and issues for Congress. CRS Report R42555. https://www.everycrsreport.com/reports/R42555.html

Alappat. 1994. In re Alappat. No. 92-1381. US Federal Circuit Court, July 29. https://law.justia.com/cases/federal/appellate-courts/F3/33/1526/513542/

Albergotti, Reed 2024. Amazon announces new "Rainier" AI compute cluster with Anthropic. Semafor. https://www.semafor.com/article/12/03/2024/amazon-announces-new-rainier-ai-compute-cluster-with-anthropic

Albertus, Michael, and Victor Menaldo. 2018. Authoritarianism and the Elite Origins of Democracy. Cambridge: Cambridge University Press.

Alcatel-Lucent, Ericsson, NEC, NextG Networks, Nokia, Nokia Siemens Networks, Nortel, and Sony Ericsson. 2008. Joint Proposal on 4G Mobile Broadband: Standard Essential Patents. ETSI.

Alcoa. 1945. United States v. Aluminum Co. of America, No. 144. US Court of Appeals, Second Circuit, March 12. https://law.justia.com/cases/federal/appellate-courts/F2/148/416/1503668/

Alhassani, Mehdi, and Anthony Bak. 2025. America Is Winning the Wrong AI Race. The Wall Street Journal. https://www.wsj.com/articles/america-is-winning-the-wrong-ai-race–

Alice Corp. 2014. Alice Corp. v. CLS Bank International. No. 13-298. US Supreme Court, June 19. https://supreme.justia.com/cases/federal/us/573/208

Alicke, Knut, Tacy Foster, Katharina Hauck, and Vera Trautwein. 2023. "Tech and regionalization bolster supply chains, but complacency looms." McKinsey & Company.

Aljunaid, Saif Khalifa, Almheiri, Saif Jasim, Dawood, Hussain, and Khan, Muhammad Adnan. (2025). Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection. Journal of Risk and Financial Management 18(4): 179.

Allen, Robert. 2009. The British Industrial Revolution in Global Perspective. Cambridge: Cambridge University Press.

Allied Orthopedic Appliances. 2010. Allied Orthopedic Appliances Inc. v. Tyco Health Care Group LP, No. 08-56314, 9th Cir., Jan. 6. https://law.justia.com/cases/federal/appellate-courts/F3/592/991

Allison, John, and Emerson Tiller. 2003. "The Business Method Patent Myth." Berkeley Technology Law Journal 18: 987-1084.

Alok, Yashashwy 2024. Top 10 Open-Source AI Libraries for Developers. https://www.analyticsvidhya.com/blog/2024/12/open-source-ai-libraries

Alphabet Inc. 2020–2025. Annual Reports and SEC Filings. SEC.

Alphabet Inc. 2025. “2024 Q4 Earnings Call” (transcript). Alphabet Investor Relations. https://abc.xyz/investor/events/event-details/2025/2024-Q4-Earnings-Call/

Alsop, Thomas. 2025. Qualcomm R&D expenditures worldwide 2016–2024. Statista. https://www.statista.com/statistics/980315/research-and-development-expenditures-of-qualcomm/

Alston, Eric, Lee Alston, Bernardo Mueller, and Tomas Nonnenmacher. 2018. Institutional and Organizational Analysis. Cambridge: Cambridge University Press.

Altman, Sam. 2025. “Reflections.” January 5, 2025. https://blog.samaltman.com/reflections

Alvarez, Robert. 2025. "How to Accelerate Training for Large AI Models with Pure Storage FlashBlade and PyTorch Checkpointing." Purely Technical (blog), Pure Storage. https://blog.purestorage.com/purely-technical/pure-storage-flashblade-and-pytorch-asynchronous-checkpointing/

Amazon Inc. 2024. "AWS Trainium2 Instances Now Generally Available." US Press Center, December 03. https://press.aboutamazon.com/2024/12/aws-trainium2-instances-now-generally-available

Amazon Inc. 2025. “Amazon.com (AMZN) Q4 2024 Earnings Call Transcript.” The Motley Fool. https://www.fool.com/earnings/call-transcripts/2025/02/06/amazoncom-amzn-q4-2024-earnings-call-transcript/

Amazon Staff. 2024. Amazon and Anthropic deepen their shared commitment to advancing generative AI. Amazon News. https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment

AMD Inc. 2025. Current Report (Form 8-K), “Entry into a Material Definitive Agreement” (Warrant to OpenAI OpCo, LLC). SEC. https://ir.amd.com/financial-information/sec-filings/content/0001193125-25-230895/0001193125-25-230895.pdf

American Can. 1949. United States v. American Can Co., No. 26345-H. US District Court, N.D. California, S.D., November 10. https://law.justia.com/cases/federal/district-courts/FSupp/87/18/

American Express. 2018. Ohio v. American Express Co., 585 U.S. ___ (138 S. Ct. 2274). https://supreme.justia.com/cases/federal/us/585/16-1454

AIPLA. 2019. Report of the Economic Survey. AIPLA. https://www.aipla.org/detail/journal-issue/2019-report-of-the-economic-survey

American Nuclear Society 2025. Talen and Amazon expand their partnership for Pennsylvania nuclear. American Nuclear Society Newswire. https://www.ans.org/news/2025-06-16/article-7113/talen-and-amazon-expand-their-partnership-for-pennsylvania/

Aminabadi, Reza Yazdani, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. 2022. DeepSpeed Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. arXiv preprint arXiv: 2207.00032. https://arxiv.org/abs/2207.00032

Anderson. 2024. Anderson v. TikTok, Inc. No. 22-3061. US Court of Appeals, Third Circuit, August 27. https://law.justia.com/cases/federal/appellate-courts/ca3/22-3061/22-3061-2024-08-27.html

Andreessen, Marc. 2011. "Why Software Is Eating the World." The Wall Street Journal. https://www.wsj.com/articles/SB10001424053111903480904576512250915629460

Anthropic. 2024. Introducing the Model Context Protocol. Anthropic. https://www.anthropic.com/news/model-context-protocol

Anthropic. 2025a. Piloting Claude for Chrome. Anthropic. https://www.anthropic.com/news/claude-for-chrome?_bhlid=bd9fae9ffe489cd5db411d4be7bf4c89adf41d9c

Anthropic. 2025c. “Expanding our use of Google Cloud TPUs and Services.” Anthropic News. https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

Anthropic. 2025d. "Anthropic invests $50 billion in American AI infrastructure." Anthropic News. https://www.anthropic.com/news/anthropic-invests-50-billion-in-american-ai-infrastructure

Anthropic. 2025e. Anthropic Economic Index: Insights from Claude 3.7 Sonnet. https://www.anthropic.com/news/anthropic-economic-index-insights-from-claude-sonnet-3-7

Appen. 2021. Overcoming AI Deployment Challenges/5 Tips for Handling AI Obstacles. Appen. https://www.appen.com/blog/overcoming-ai-deployment-challenges

Apple Computer. 1994. Apple Computer, Inc. v. Microsoft Corp. Nos. 93-16867, 93-16869 and 93-16883. US Court of Appeals, Ninth Circuit, September 19. https://law.justia.com/cases/federal/appellate-courts/F3/35/1435/605245/

Apple Developer. n.d.a. “Get Started — iOS.” Apple Developer Documentation. https://developer.apple.com/ios/get-started

Apple Developer. n.d.b. "Core Location." Apple Developer Documentation. https://developer.apple.com/documentation/corelocation/

Apple Developer. n.d.c. “Configuring your app to use location services.” Apple Developer Documentation. https://developer.apple.com/documentation/corelocation/configuring-your-app-to-use-location-services

Apple Developer. n.d.d. “In-app purchase.” Apple Developer Documentation. https://developer.apple.com/design/human-interface-guidelines/in-app-purchase

Apple Developer. n.d.e. “Handling location updates in the background.” Apple Developer Documentation. https://developers.apple.com/documentation/corelocation/handling-location-updates-in-the-background

Apple Developer. n.d.f. “Getting the current location of a device.” Apple Developer Documentation. https://developer.apple.com/documentation/corelocation/getting-the-current-location-of-a-device

Apple Inc. 2019b. Apple Inc. v. Robert Pepper, et al., No. 17-204, U.S. Supreme Court, May 13. https://scholar.google.com/scholar_case?case=15157888566008153651&q=apple+v+pepper&hl=en&as_sdt=6,48&as_vis=1

Apple Inc. 2020–2025. Annual Reports and SEC Filings. SEC.

Apple Inc. 2024a. Annual Report 2024 (Form 10-K). Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/320193/000032019324000123/aapl-20240928.htm

Apple Inc. 2023b. “Apple unveils iPhone 15 Pro and iPhone 15 Pro Max.” Press Release. Apple Newsroom. https://www.apple.com/newsroom/2023/09/apple-unveils-iphone-15-pro-and-iphone-15-pro-max/

Apple Inc. 2024b. "iPhone 16 – Technical Specifications." https://www.apple.com/iphone-16/specs/

Apple Inc. 2024c. “Introducing Apple Intelligence for iPhone, iPad, and Mac.” Apple Newsroom. https://www.apple.com/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/

Apple Support. n.d. “Use Continuity to work across Apple devices.” Apple Support. https://support.apple.com/guide/iphone/intro-to-continuity-iphf5fa30b66/ios

Apple. 2021a. "A Day in the Life of Your Data." https://www.apple.com/privacy/docs/A_Day_in_the_Life_of_Your_Data.pdf

Apple. 2021b. "App Tracking Transparency." Apple Developer Documentation. https://developer.apple.com/documentation/apptrackingtransparency

Apple. 2021c. “Data Privacy Day at Apple.” Apple Newsroom. https://www.apple.com/newsroom/2021/01/data-privacy-day-at-apple-improving-transparency-and-empowering-users

Apple. 2022. “Deploying Transformers on the Apple Neural Engine.” Apple Machine Learning Research. https://machinelearning.apple.com/research/neural-engine-transformers

Apple. 2024a. Apple Platform Security. https://help.apple.com/pdf/security/en_US/apple-platform-security-guide.pdf

Apple. 2024b. “Introducing Apple’s On-Device and Server Foundation Models.” Apple Machine Learning Research. https://machinelearning.apple.com/research/introducing-apple-foundation-models

Apple. 2024c. “Private Cloud Compute: A new frontier for AI privacy in the cloud.” Apple Security Research. https://security.apple.com/blog/private-cloud-compute

Apple. 2024d. “Apple Intelligence Foundation Language Models.” arXiv preprint: arXiv 2407.21075. https://arxiv.org/pdf/2407.21075

Apple. 2025. “Apple Intelligence Foundation Language Models: Tech Report 2025.” arXiv preprint: arXiv 2507.13575. https://arxiv.org/pdf/2507.13575

Applegate, Lynda, and Kevin Davis. 1995. “Xerox: Outsourcing Global Information Technology Resources.” Harvard Business School Case 9-195-158, Harvard Business School Publishing. Available at: https://hbsp.harvard.edu/product/195158-PDF-ENG

Apptronik. 2024. Apptronik partners with Google DeepMind robotics to accelerate advancement on AI-powered humanoid robots. Apptronik Press Release. https://apptronik.com/news-collection/apptronik-partners-with-google-deepmind-robotics

Areeda, Phillip, and Herbert Hovenkamp. 1978. Antitrust Law: An Analysis of Antitrust Principles and Their Application. New York: Aspen Publishers.

Areeda, Phillip, and Herbert Hovenkamp. 2003. Antitrust Law: An Analysis of Antitrust Principles and Their Application (2nd ed.). New York: Aspen Publishers.

Areeda, Phillip, and Herbert Hovenkamp. 2023. Antitrust Law: An Analysis of Antitrust Principles and Their Application (5th ed.). New York: Wolters Kluwer.

Areeda, Phillip, and Donald Turner. 1975. "Predatory Pricing and Related Practices Under Section 2 of the Sherman Act." Harvard Law Review 88(4): 697–733.

Areeda, Phillip, and Donald Turner. 1978. Antitrust Law. Boston: Little, Brown.

Areeda, Phillip, and Herbert Hovenkamp. 2020. Antitrust Law. 5th ed. New York: Wolters Kluwer.

Areeda, Phillip. 1989. "Essential Facilities." Antitrust Law Journal 58(3): 841–53.

Argenton, Cédric, and Jens Prüfer. 2012. "Search Engine Competition with Network Externalities." Journal of Competition Law and Economics 8(1): 73-105.

Aridor, Guy, Yeon-Koo Che, Brett Hollenbeck, Maximilian Kaiser, and Daniel McCarthy. 2025. “Evaluating the Impact of Privacy Regulation on E-Commerce Firms.” Working paper. https://www.anderson.ucla.edu/sites/default/files/document/2025-05/att_privacy.pdf

Arista Networks. 2024. Arista Unveils Etherlink AI Networking Platforms. Arista Networks Press Release. https://investors.arista.com/Communications/Press-Releases-and-Events/Press-Release-Detail/2024/Arista-Unveils-Etherlink-AI-Networking-Platforms/default.aspx

Arm Holdings plc. 2023. Registration Statement on Form F-1. U.S. Securities and Exchange Commission. https://investors.arm.com/node/6921/html

Armbrust, Michael, Ali Ghodsi, Reynold Xin, and Matei Zaharia. 2021. "Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics." Proceedings of CIDR 2021.

Armstrong, Mark. 2006. "Competition in Two-Sided Markets." The RAND Journal of Economics 37(3): 668–91.

Arnold, Thurman Wesley 1941. The Bottlenecks of Business. New York: Reynal and Hitchcock.

Arora, Ashish, Andrea Fosfuri, and Alfonso Gambardella. 2001. Markets for Technology. Cambridge, MA: MIT Press.

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources for Invention." In The Rate and Direction of Inventive Activity: Economic and Social Factors, edited by Richard Nelson. Princeton, NJ: Princeton University Press, 609–25.

Arroyo, Carmen, and Edward Ludlow. 2025. “xAI to Raise $20 Billion After Nvidia and Others Boost Round.” Bloomberg. https://www.bloomberg.com/news/articles/2025-10-07/musk-s-xai-nears-20-billion-capital-raise-tied-to-nvidia-chips?embedded-checkout=true

ASME. 2025. “Special Issue: Human–Robot Collaboration in Industry 5.0.” https://asmejcise.org/2025/04/07/special-issue-human-robot-collaboration-in-industry-5-0/

ASML. 2024. ASML Annual Report 2024. ASML Holding N.V. Investor Relations. https://www.asml.com/en/investors/annual-report/2024

ASML. 2025. “EUV lithography.” ASML Products – EUV lithography systems. https://www.asml.com/en/products/euv-lithography-systems

ASPDAC. 2025. “Chiplet Package Design: Complexity, Reliability, and Advanced Packaging.” https://www.aspdac.com/aspdac2025/archive/pdf/7F-3.pdf

ASTC. 2025. "New ASTC Survey Shows Gaps Between Public Support for Science and Understanding of Federal Actions." ASTC News. https://www.astc.org/astc-news-announcements/new-astc-survey-shows-gaps-between-public-support-for-science-and-understanding-of-how-science-is-impacted-by-federal-actions/

AT&T 2018. United States v. AT&T Inc., 310 F. Supp. 3d 161 (D.D.C.). https://law.justia.com/cases/federal/district-courts/district-of-columbia/dcdce/1:2017cv02511/182300/172

AT&T 2019. United States v. AT&T Inc., No. 18-5214, 916 F.3d 1029 (D.C. Cir.). https://law.justia.com/cases/federal/appellate-courts/cadc/18-5214/18-5214-2019-02-26.html

AT&T 1999. AT&T Corp. v. Excel Communications, Inc., Excel Communications Marketing, Inc., and Excel Telecommunications, Inc. No. 98-1338. US Court of Appeals, Federal Circuit, April 14. https://law.justia.com/cases/federal/appellate-courts/F3/172/1352/599511/

Atkinson, Robert, Daniel Castro, and Stephen Ezell. 2008. "The Digital Road to Recovery: A Stimulus Plan to Create Jobs, Boost Productivity and Revitalize America." Information Technology and Innovation Foundation. https://d1bcsfjk95uj19.cloudfront.net/files/Digital_Road_to_Recovery.pdf

Atkinson, Robert. 2025. “No, AI Robots Won’t Take All Our Jobs: Instead, they will boost productivity, lower prices and spur the evolution of the labor market.” The Wall Street Journal.

https://www.wsj.com/opinion/no-ai-robots-wont-take-all-our-jobs-employment-productivity-innovation-41068792

ATOM Project Substack. 2025. One Small Step for ATOM, one giant leap for OpenAI. https://atomproject.substack.com/p/one-small-step-for-atom-one-giant

Austin, Steven. 2025. How many people work at Anthropic (Claude)? Comprehensive workforce analysis for 2025. Marketingscoop. https://www.marketingscoop.com/website/seo/how-many-people-work-at-anthropic-claude-comprehensive-workforce-analysis-for-2025/

Authers, John. 2026. Greenland Isn't as Big as it Looks in Risk. Points of Return. Bloomberg. https://www.bloomberg.com/opinion/newsletters/2026-01-20/greenland-isn-t-as-big-as-it-looks-in-risk?srnd=phx-opinion

Autor, David, David Dorn, Lawrence Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." Quarterly Journal of Economics 135(2): 645–709.

Autor, David, and Michael Handel. 2013. "Putting Tasks to the Test: Human Capital, Job Tasks, and Wages." Journal of Labor Economics 31(1): S59–S96.

Autor, David. 2015. "Why Are There Still So Many Jobs? The History and Future of Workplace Automation." Journal of Economic Perspectives 29(3): 3–30.

Autor, David. 2024. "Applying AI to Rebuild Middle Class Jobs." NBER Working Paper No. 32140. Cambridge, MA: National Bureau of Economic Research.

AWS. 2024a. High Performance Computing on AWS. AWS. https://aws.amazon.com/hpc

AWS. 2024b. Generative AI inference architecture and best practices on AWS (AWS inference stack). Available at: <https://docs.aws.amazon.com/prescriptive-guidance/latest/gen-ai-inference-architecture-and-best-practices-on-aws/aws-inference-stack.html

AWS. 2024c. “Anthropic's Claude 3 Opus model now available on Amazon Bedrock.” AWS What’s New. https://aws.amazon.com/about-aws/whats-new/2024/04/anthropics-claude-3-opus-amazon-bedrock

AWS. 2025. Amazon Bedrock Documentation. AWS. https://docs.aws.amazon.com/bedrock

Ayodele, Abiola. 2025. OSAT Semiconductor Services: The Backbone of Outsourced Chip Assembly and Testing. Wevolver. https://www.wevolver.com/article/osat-semiconductor-services-the-backbone-of-outsourced-chip-assembly-testing

Azure AI Studio. 2025. Endpoints for inference in production. Azure AI Studio. https://learn.microsoft.com/en-us/azure/machine-learning/concept-endpoints

Baack, Stefan. 2024. "A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl." Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3630106.3659033

Backus, Matthew, Christopher Conlon, and Michael Sinkinson. 2021. "Common Ownership in America: 1980–2017." American Economic Journal: Microeconomics 13(3): 273–308.

Baek, Jinheon, Nirupama Chandrasekaran, Silviu Cucerzan, and Allen Herring. 2024. Knowledge-Augmented Large Language Models for Personalized Contextual Query Suggestion. Proceedings of the ACM Web Conference (WWW ’24). https://arxiv.org/pdf/2311.06318

Bahar, Dany. 2018. “The Middle Productivity Trap: Dynamics of Productivity Dispersion.” Economics Letters 167: 60–66.

Bailey, Martha, John DiNardo, and Bryan Stuart. 2021. "The Economic Impact of a High National Minimum Wage: Evidence from the 1966 Fair Labor Standards Act." Journal of Labor Economics, Vol. 39(S2): S329-S367.

Bain, Joe. 1956. Barriers to New Competition. Cambridge, MA: Harvard University Press.

Baker, Jonathan. 1997. The Problem with Baker Hughes and Syufy: On the Role of Entry in Merger Analysis. Antitrust Law Journal 65(2): 353-374.

Baker, Jonathan and Carl Shapiro. 2010. Reinvigorating Horizontal Merger Enforcement. In Robert Pitofsky (Ed.), How the Chicago School Overshot the Mark: The Effect of Conservative Economic Analysis on U.S. Antitrust New York: Oxford University Press, 235-291.

Baker, Jonathan. 2007. "Market Definition." Antitrust Law Journal 74(1): 129–173.

Bakos, Yannis. 1997. "Reducing Buyer Search Costs: Implications for Electronic Marketplaces." Management Science 43(12): 1676–1692.

Baldwin, Richard. 2016. The Great Convergence. Cambridge, MA: Harvard University Press.

Ball Memorial Hospital. 1986. Ball Memorial Hospital, Inc. v. Mutual Hospital Insurance, Inc., No. 85-1481. US Court of Appeals, Seventh Circuit, March 4. https://law.justia.com/cases/federal/appellate-courts/F2/784/1322

Balp, Laura. 2025. Storage Next: Why AI Workloads Need a Kitchen Built for Speed. ScaleFlux Blog. https://scaleflux.com/storage/storage-next-serving-data-like-a-world-class-restaurant

Balser, Jimmy, Minon Schwartz and Jon Shimabukuro. 2025. "A New Civil Service 'Policy/Career' Schedule." CRS Report LSB11262. https://www.congress.gov/crs-product/LSB11262

Bandyopadhyay, Subhayu, Maximiliano Dvorkin, Fernando Leibovici, and Ana Maria Santacreu. 2022. The Shifting Tides of Global Trade. Federal Reserve Bank of St. Louis. https://www.stlouisfed.org/annual-report/2022/shifting-tides-global-trade.

Banga, Rashmi. 2013. Measuring value in global value chains. UNCTAD. https://unctad.org/system/files/official-document/ecidc2013misc1_bp8.pdf

Bantock, Asher 2025. The SignalFire State of Tech Talent Report – 2025. SignalFire. https://www.signalfire.com/blog/signalfire-state-of-talent-report-2025

Barath Harithas, Kyle Meng, Evan Brown, and Catharine Mouradian. 2025. The Liberation Day Tariffs. Washington, DC: Center for Strategic and International Studies. https://www.csis.org/analysis/liberation-day-tariffs-explained

Barenholtz, Lior. 2025. ChatGPT monthly visits: SEO analysis and competitor comparisons. Similarweb. https://www.similarweb.com/blog/marketing/geo/chatgpt-monthly-visits/

Barkai, Simcha. 2020. “Declining Labor and Capital Shares.” Journal of Finance 75 (5): 2421–2463.

Barnes and Noble. 2014. Nguyen v. Barnes and Noble Inc., No. 12-56628, U.S. Court of Appeals, Ninth Circuit, August 16. https://scholar.google.com/scholar_case?case=11003811139217543321

Barnes. 2009. Barnes v. Yahoo!, Inc. No. 06-56662. US Court of Appeals, Ninth Circuit, June 22. https://law.justia.com/cases/federal/appellate-courts/F3/570/1096

Barnett, Jonathan. 2020. Innovators, Firms, and Markets: The Organizational Logic of Intellectual Property. New York: Oxford University Press.

Barnett, Jonathan. 2017. "Has the Academy Led Patent Law Astray?" Berkeley Technology Law Journal 32(4): 1313–1386.

Barron, David. 2010. Best Practices for OLC Legal Advice and Written Opinions. Memorandum for Attorneys of the Office. DoJ, Office of Legal Counsel. https://www.justice.gov/olc/page/file/1511836/dl

Baroudy, Kim, Jonatan Janmark, Tobias Strålin, Abhi Satyavarapu, and Zeno Ziemke. 2020. “Europe’s start-up ecosystem: Heating up, but still facing challenges.” McKinsey & Company.

https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/europes-start-up-ecosystem-heating-up-but-still-facing-challenges

Barroso, Luiz André, Jeffrey Dean, and Urs Holzle. 2003. Web search for a planet: The Google cluster architecture. IEEE micro, 23(2): 22-28.

Barry Wright. 1983. Barry Wright Corp. v. ITT Grinnell Corp., No. 83-1292. US Court of Appeals, First Circuit, December 29. https://law.justia.com/cases/federal/appellate-courts/F2/724/227/265312/

Bartz et al. 2025. Bartz et al. v. Anthropic PBC, No. 3:24-cv-05417. US District Court for the Northern District of California, August 19, 2024. https://docs.justia.com/cases/federal/district-courts/california/candce/3:2024cv05417/434709/231

Battelle, John. 2005. The Search. New York: Portfolio.

Baumol, William. 1982. "Contestable Markets: An Uprising in the Theory of Industry Structure." The American Economic Review 72(1): 1–15.

Harrison, Joseph, Matthew Josefy, Matias Kalm, and Ryan Krause. 2023. "Using supervised machine learning to scale human-coded data: A method and dataset in the board leadership context." Strategic Management Journal 44(7): 1780-1802.

Bayard, Thomas, and Kimberly Ann Elliott. 1994. Reciprocity and Retaliation in U.S. Trade Policy. Washington, DC: Institute for International Economics.

BEA. 2024. Trade in Services in 2023. BEA. https://apps.bea.gov/scb/issues/2024/10-october/1024-international-services.htm

BEA. 2025. U.S. International Trade in Goods and Services, December and Annual 2024. https://www.bea.gov/news/2025/us-international-trade-goods-and-services-december-and-annual-2024

Beauchene, Vinciane, Sylvain Duranton, Nipun Kalra, and David Martin. 2025. “AI at Work 2025: Momentum Builds, but Gaps Remain.” Boston Consulting Group, Slideshow. https://web-assets.bcg.com/fd/0d/bcc5dfae4cbaa08c718b95b16cf5/ai-at-work-2025-slideshow-june-2025-edit-02.pdf

Bell, Emily, and Taylor Owen. 2017. "The Platform Press." Tow Center for Digital Journalism, Columbia Journalism School. https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php

Bendavid, Naftali, and Abha Bhattarai. 2025. “Trump’s Risky Role as the Would-Be CEO of America.” The Washington Post. https://www.washingtonpost.com/politics/2025/08/12/trump-corporations-ceo-economy-tariffs/

Benesch, Friedlander, Coplan and Aronoff LLP. 2008. “Qualcomm Highlights the Need for In-House and Outside Counsel to Work Closely to Conduct E-Discovery.” E-Discovery Bulletin. https://www.beneschlaw.com/a/web/687/IP_eDoc_Qualcomm.pdf

Bengio, Yoshua, Sasha Luccioni, Edward Raff, Abhishek Gupta, Daniel McDuff, David Rolnick, Emma Strubell, Irene Solaiman, Margaret Mitchell, and Timnit Gebru. 2023. “Collective Governance of Open Foundation Models.” arXiv preprint: arXiv.2307.08774. https://doi.org/10.48550/arXiv.2307.08774

Benjamin, Dan, and Matthews, Kent. 1992. U.S. and U.K. Unemployment Between the Wars: A Doleful Story. Institute for Economic Affairs, London.

Benston, George. 1990. The Separation of Commercial and Investment Banking. New York: Oxford University Press.

Beraja, Martin, Andrew Kao, David Yang, and Noam Yuchtman. 2023. "AI-tocracy." Quarterly Journal of Economics 138(3): 1349–1402.

Bergen, Mark, and Jennifer Surane. 2017. "Google Will Stop Reading Your Emails for Gmail Ads." Bloomberg. https://www.bloomberg.com/news/articles/2017-06-23/google-will-stop-reading-your-emails-for-gmail-ads

Berkey Photo. 1979. Berkey Photo, Inc. v. Eastman Kodak Co., Nos. 1019, 1070. US Court of Appeals, Second Circuit, June 25. https://law.justia.com/cases/federal/appellate-courts/F2/603/263/105215

Bessen, James, and Michael Meurer. 2008. Patent Failure: How Judges, Bureaucrats, and Lawyers Put Innovators at Risk. Princeton: Princeton University Press.

Bessen, James, and Robert Hunt. 2007. “An Empirical Look at Software Patents.” Journal of Economics and Management Strategy 16(1): 157–189.

Bessen, James, and Stefano Calligaris. 2018. “Proprietary IT Intensity and the Persistence of Superstar Returns.” NBER Working Paper No. 24259. Cambridge, MA: National Bureau of Economic Research.

Bessen, James. 2015. Learning by Doing: The Real Connection between Innovation, Wages, and Wealth. New Haven, CT: Yale University Press.

Bessen, James. 2020. Industry Concentration and Information Technology. Journal of Law and Economics 63(3): 531–570.

Bick, Alexander, Adam Blandin, and David Deming. 2025. The Rapid Adoption of Generative AI, Federal Reserve Bank of St. Louis Working Paper 2024-027. https://doi.org/10.20955/wp.2024.027

BigCommerce. (n.d.). What is a Facebook Custom Audience and how can they grow online stores? BigCommerce. https://www.bigcommerce.com/glossary/facebook-customer-audience/

Biswas, Prassana. 2025. CUDA vs SYCL: A Comparison of GPU Programming Models. LinkedIn. https://www.linkedin.com/pulse/cuda-vs-sycl-comparison-gpu-programming-models-prasanna-biswas-pdwsc

BitLaw. 2025. “History of Software Patents, from Benson, Flook, and Diehr to Bilski and Alice.” Beck and Tysver. https://www.bitlaw.com/software-patent/history.html.

Bjorhovde, Runar. 2024. "How much money do smartphone vendors actually earn per device?" LinkedIn.

Blackstone. (2024, May 17). CoreWeave Secures $7.5 Billion Debt Financing Facility led by Blackstone and Magnetar. https://www.blackstone.com/news/press/coreweave-secures-7-5-billion-debt-financing-facility-led-by-blackstone-and-magnetar/

Blevins, Emily, Alice Grossman, and Karen Sutter. 2023. CHIPS Act of 2022. CRS Report R47523. https://www.congress.gov/crs-product/R47523

Blind, Knut. 2004. The Economics of Standards: Theory, Evidence, Policy. Cheltenham, England: Edward Elgar Publishing.

Block, Fred 2008. Swimming Against the Current: The Rise of a Hidden Developmental State in the United States. Politics and Society, 36(2): 169–206.

Block, Fred, and Matthew Keller. 2011. State of Innovation: The U.S. Government’s Role in Technology Development. New York: Routledge.

Block. 2025. "Block, Anthropic, and OpenAI Launch the Agentic AI Foundation." Block Inside. https://block.xyz/inside/block-anthropic-and-openai-launch-the-agentic-ai-foundation

Bloom, Benjamin 1984. “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring.” Educational Researcher 13(6): 4–16.

Bloomberg News. 2025. “Klarna Turns From AI to Real-Person Customer Service.” Bloomberg. https://www.bloomberg.com/news/articles/2025-05-08/klarna-turns-from-ai-to-real-person-customer-service

BLS. 2011. Handbook of Methods. DoL. https://www.bls.gov/opub/hom/pdf/inp-20110509.pdf

BLS. 2025a. “Year-over-Year Inflation Rate: CPI-U (All Items), Annual Average, 2002–2024.” DoL. https://www.bls.gov/pir/spm/spm_chart5_2024data.htm

BLS 2025b. Productivity and Costs: Fourth Quarter and Annual Averages 2024, Preliminary. DoL. https://www.bls.gov/news.release/archives/prod2_02062025.htm

BLS. 2025c. Productivity and Costs: Second Quarter 2025, Revised. DoL. https://www.bls.gov/news.release/prod2.nr0.htm

BLS. 2025d. “Office of Productivity and Technology: Concepts.” BLS Handbook of Methods. DoL. https://www.bls.gov/opub/hom/opt/concepts.htm

BLS. 2025e. “Office of Productivity and Technology: Data Sources.” BLS Handbook of Methods. DoL.. https://www.bls.gov/opub/hom/opt/data.htm

BLS. n.d. Consumer Price Index (CPI). DoL. https://www.bls.gov/cpi/

Bohn, Robert, Christopher Greer, and Jason Kahn. 2023. Report of the Advanced Communications Technologies Working Group (NIST Interagency Report 8483). NIST. https://tsapps.nist.gov/publication/get_pdf.cfm

Bonacich, Edna, and Khaleelah Hardie. 2006. “Wal-Mart and the Logistics Revolution.” In Wal-Mart: The Face of Twenty-First-Century Capitalism, edited by Nelson Lichtenstein, New York: The New Press, 163–187.

Bonatesta, Lara. 2025. “Forty-seven states have enacted deepfake legislation since 2019.” Ballotpedia News. https://news.ballotpedia.org/2025/07/22/forty-seven-states-have-enacted-deepfake-legislation-since-2019/

Bonnet, Alexandre. 2024. ONNX Standardized Format: The Universal Translator for AI Models. Encord. https://encord.com/blog/onnx-open-neural-network-exchange-format/

Borin, Alessandro, and Michele Mancini. 2021. “Supply Chain Fragmentation and the Global Trade Elasticity: A New Accounting Framework.” IMF Economic Review 69 (4): 656–680.

Bork, Robert. 1978. The Antitrust Paradox. New York: Basic Books.

Borkar, Shekhar, Saurabh Dighe, and Nirav Kahn. 2024. "Accelerating Industry-Wide Innovations in Datacenter Infrastructure and Security." Microsoft Azure Blog. https://azure.microsoft.com/en-us/blog/accelerating-industry-wide-innovations-in-datacenter-infrastructure-and-security/

Bort, Julie. 2025a. "Google Launched Its Deepest AI Research Agent Yet—on the Same Day OpenAI Dropped GPT-5.2." TechCrunch. https://techcrunch.com/2025/12/11/google-launched-its-deepest-ai-research-agent-yet-on-the-same-day-openai-dropped-gpt-5-2/

Bort, Julie. 2025b. "In another chess move with Microsoft, OpenAI is pouring $12B into CoreWeave." https://techcrunch.com/2025/03/10/in-another-chess-move-with-microsoft-openai-is-pouring-12b-into-coreweave/

Boudreau, Kevin J., and Hagiu, Andrei 2009. Platform rules: Multi-sided platforms as regulators." *Platforms, markets and innovation* 1(0): 163-191.

Bousquette, Isabelle, Belle Lin, Tom Loftus, and Steven Rosenbush. 2025. "What Are Companies Actually Doing With AI? Our Reporters Talk It Out." Wall Street Journal.

https://www.wsj.com/articles/what-are-companies-actually-doing-with-ai-our-reporters-talk-it-out-a12dd305

Bowman, Ward. 1957. Tying Arrangements and the Leverage Problem. Yale Law Journal, 67(1): 19-36.

Boyle, Alan, Taylor Soper, and Todd Bishop. 2020. "Apple acquires Xnor.ai, edge AI spin-out from Paul Allen's AI2, for price in $200M range." GeekWire. https://www.geekwire.com/2019/ai2-china-us-research/

Bradford, Anu. 2020. The Brussels Effect. New York: Oxford University Press.

Bradford, Anu. 2023. Digital Empires. Oxford: Oxford University Press.

Brantle, Thomas, and M. Hosein Fallah. 2007. “Complex Innovation Networks, Patent Citations and Power Laws.” In PICMET ’07: Portland International Conference on Management of Engineering and Technology, August 5–9, Portland, OR. IEEE. https://doi.org/10.1109/PICMET.2007.4349390

Bratton, Laura. 2025. “Big Tech’s AI Investments Set to Spike to $364 Billion in 2025.” Yahoo Finance. https://finance.yahoo.com/news/big-techs-ai-investments-set-to-spike-to-364-billion-in-2025-as-bubble-fears-ease-143203885.html

Braun, Alan. 2025. Scaling responsibly: evolving our API rate limits to power the next generation of Atlassian Cloud. Atlassian Blog. https://www.atlassian.com/blog/platform/evolving-api-rate-limits

Bresnahan, Timothy, and Manuel Trajtenberg. 1995. "General Purpose Technologies: 'Engines of Growth'?" Journal of Econometrics 65(1): 83–108.

Bresnahan, Timothy, and Shane Greenstein. 1996. Technical Progress and Co-Invention in Computing and in the Uses of Computers. Brookings Papers on Economic Activity:

Microeconomics: 1–77. https://www.brookings.edu/wp-content/uploads/1996/01/1996_bpeamicro_bresnahan.pdf

Bresnahan, Timothy, Erik Brynjolfsson, and Lorin Hitt. 2002. "Information Technology, Workplace Organization, and the Demand for Skilled Labor." The Quarterly Journal of Economics 117(1): 339-376.

Bresnahan, Timothy. 2010. General Purpose Technologies. In Handbook of the Economics of Innovation, Vol. 2, edited by Bronwyn Hall and Nathan Rosenberg. Amsterdam, Holland: Elsevier, 761-791.

Breyer, Stephen, Richard Stewart, Cass Sunstein, Adrian Vermeule, and Michael Herz. 2022. Administrative Law and Regulatory Policy: Problems, Text, and Cases. 9th ed. Burlington, MA: Aspen Publishing.

Bria, Francesca, Paul Timmers and Fausto Gernone. 2025. EuroStack – A European Alternative for Digital Sovereignty. Bertelsmann Stiftung. https://www.bertelsmann-stiftung.de/en/our-projects/reframetech-algorithmen-fuers-gemeinwohl/project-news/eurostack-a-european-alternative-for-digital-sovereignty

Brock, William, and Steven Durlauf. 2001. "Discrete Choice with Social Interactions." Review of Economic Studies 68(2): 235-260.

Brooke Group Ltd. 1993. Brooke Group Ltd. v. Brown and Williamson Tobacco Corp., No. 92-466. U.S. Supreme Court, June 21. https://supreme.justia.com/cases/federal/us/509/209

Brougher, Jim. 2025. “Reddit's New AI Licensing Deal Shows How Content Co.s Get Paid to Train LLMs.” Media + the Machine (Substack), September 29, 2025. https://mediaandthemachine.substack.com/p/reddits-new-ai-licensing-deal-shows.

Brown Shoe. 1962. Brown Shoe Co. v. United States, No. 4. US Supreme Court, June 25. https://supreme.justia.com/cases/federal/us/370/294/

Brown, A. B. 2020. Record 2.7M industrial robots are in operation, but demand has slowed. Supply Chain Dive.

Brynjolfsson, Erik. 1993. The Productivity Paradox of Information Technology. Communications of the ACM 36(12): 66-77.

Brynjolfsson, Erik, Avinash Collis, W. Erwin Diewert, Felix Eggers, and Kevin Fox. 2019. GDP-B: Accounting for the value of new and free goods in the digital economy. NBER Working Paper No. 25695. Cambridge, MA: National Bureau of Economic Research.

Brynjolfsson, Erik, and Lorin Hitt. 2000. "Beyond Computation." Journal of Economic Perspectives 14(4): 23–48.

Brynjolfsson, Erik, Bharat Chandar, and Ruyu Chen. 2025. Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence. Stanford Digital Economy Lab Working Paper. https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2021. The Productivity J-Curve. American Economic Journal: Macroeconomics 13(1): 333–372.

Buchanan, James, and Gordon Tullock. 1962. The Calculus of Consent: Logical Foundations of Constitutional Democracy. Ann Arbor: University of Michigan Press.

Buiten, Miriam. 2019. "Towards intelligent regulation of artificial intelligence." *European Journal of Risk Regulation* 10(1): 41-59.

BEA. 2021. "Gross Domestic Product by Industry." DoC. https://www.bea.gov/news/2022/gross-domestic-product-fourth-quarter-and-year-2021-second-estimate

Burg, David, and Jesse Ausubel 2021. Moore's Law revisited through Intel chip density. PLOS ONE 16(8): e0256245.

Burrell, Jennifer 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big Data and Society 3(1). https://journals.sagepub.com/doi/10.1177/2053951715622512

Bush, George W. 2007. Executive Order 13422: Further Amendment to Executive Order 12866 on Regulatory Planning and Review. 72 FR: 2763. https://www.federalregister.gov/documents/2007/01/23/07-293/further-amendment-to-executive-order-12866-on-regulatory-planning-and-review

Business Wire. 2025. “SambaNova Expands Deployment with SoftBank Corp. to Offer Fast AI Inference Across APAC.” Business Wire. https://www.businesswire.com/news/home/20250305799834/en/SambaNova-Expands-Deployment-with-SoftBank-Corp.-to-Offer-Fast-AI-Inference-Across-APAC/

Byrne, David, and Carol Corrado. 2017. "ICT Services and their Prices: What do they tell us about Productivity and Technology?" Finance and Economics Discussion Series 2017-015. Board of Governors of the Federal Reserve System. https://doi.org/10.17016/FEDS.2017.015

C2PA. 2023. C2PA Technical Specification, Version 2.0. C2PA. https://c2pa.org/specifications/specifications/2.0/

Caballero, Ricardo, and Adam Jaffe. 1993. “How High Are the Giants’ Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth.” In NBER Macroeconomics Annual 1993, Volume 8, edited by Olivier J. Blanchard and Stanley Fischer. Cambridge, MA: MIT Press, 15–74.

CAC. 2022. “Provisions on the Administration of Deep Synthesis of Internet-based Information Services.” Order No.12. https://www.chinalawtranslate.com/en/deep-synthesis/

CAC. 2023. “Interim Measures for the Management of Generative Artificial Intelligence Services.” Order No.15. https://www.chinalawtranslate.com/en/generative-ai-interim/

Cadwalladr, Carole, and Emma Graham-Harrison. 2018. "Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach." The Guardian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Cai, Kenrick. 2025. "Meta introduces Llama application programming interface to attract AI developers." Reuters. (Reprint: Investing.com.) https://www.investing.com/news/stock-market-news/meta-introduces-llama-application-programming-interface-to-attract-ai-developers-4011605

Caines, Colin, Sharon Jeon, and Cheyenne Quijano. 2025. "Developments in Chinese Chipmaking." FEDS Notes. U.S. Federal Reserve. https://www.federalreserve.gov/econres/notes/feds-notes/developments-in-chinese-chipmaking-20250117.html

California Dental Association. 1999. California Dental Association v. Federal Trade Commission, 526 U.S. 756. https://supreme.justia.com/cases/federal/us/526/756

Campbell-Kelly, Martin, Daniel Garcia-Swartz, Richard Lam, and Yilei Yang. 2015. "Economic and Business Perspectives on Smartphones as Multi-sided Platforms." Telecommunications Policy 39(8): 717-34.

Cantwell, Maria 2025. “Pipeline Safety Enforcement Plummets Under Trump.” U.S. Senate Committee on Commerce, Science, and Transportation.

https://www.commerce.senate.gov/2025/5/cantwell-pipeline-safety-enforcement-plummets-under-trump

Carleton, Dennis. 2007. Does antitrust need to be modernized?." Journal of Economic Perspectives 21(3): 155-176.

Carlton, Dennis W, and Jeffrey Perloff. 2005. Modern Industrial Organization, Vol. 4. Boston: Pearson/Addison Wesley.

Carr, David. 2024. ChatGPT topped 3 billion visits in September. Similarweb. https://www.similarweb.com/blog/insights/ai-news/chatgpt-topped-3-billion-visits-in-september/

Carr, Nicholas. 2003. "IT Doesn't Matter." Harvard Business Review 81(5): 41-49.

Carr, Nicholas. 2008. The Big Switch. New York: W.W. Norton.

Carter, James. 1977. Executive Order 12022—Establishing the National Commission for the Review of Antitrust Laws and Procedures. 42 FR 61441. https://www.federalregister.gov/executive-order/12022

Carter, James. 1979. Trade Agreements Program for 1979: Message to the Congress Transmitting a Report. The American Presidency Project. https://www.presidency.ucsb.edu/documents/trade-agreements-program-for-1979-message-the-congress-transmitting-report

Carter, James. 1980a. “Economic Renewal Program Remarks Announcing the Program.” The American Presidency Project at UCSB. https://www.presidency.ucsb.edu/documents/economic-renewal-program-remarks-announcing-the-program

Carter, Jimmy. 1980b. "Patent and Trademark System Reform: Statement on Signing H.R. 6933 Into Law." https://www.presidency.ucsb.edu/documents/patent-and-trademark-system-reform-statement-signing-hr-6933-into-law

Carter, James. 1980c. Executive Order 12188—International Trade Functions. 45 FR: 989. https://www.archives.gov/federal-register/codification/executive-order/12188.html

Casado, Martin, and Anne Neuberger. 2025. “China Is Winning the Race for AI Robots: Americans who want to compete have to spend more time with regulatory lawyers than engineers.” The Wall Street Journal. https://www.wsj.com/opinion/china-is-winning-the-race-for-intelligent-robots-7b2416fe

Casanova, Lourdes, and Anne Miroux. 2018. “The Competitive Edge in China’s Overseas Investments: The Golden Hand of the State.” Notes Internacionals CIDOB, no. 205 (November). Barcelona: Barcelona Centre for International Affairs.

Case, Jeffrey., Mark Fedor, Martin Lee Schoffstall, and James Davin. 1990. "A Simple Network Management Protocol (SNMP)." RFC 1157. Internet Engineering Task Force. Unpublished Paper.

Casey, Christopher. 2024. Presidential Authority to Address Tariff Barriers in Trade Agreements. CRS Report IF11400. https://www.everycrsreport.com/files/2024-10-15_IF11400_bbbd27645444fe543bee2c23cfb5d6163aaeade1.pdf

Castro, Daniel, and Michael McLaughlin. 2024. “How Innovative Is China in AI?” Information Technology and Innovation Foundation (ITIF). https://itif.org/publications/2024/08/26/how-innovative-is-china-in-ai/

Caudel, Edward, and James Magar 1986. Microcomputer system for digital signal processing (U.S. Patent No. 4,577,282). USPTO. https://image-ppubs.uspto.gov/dirsearch-public/print/downloadPdf/4577282

Cazzaniga, Mauro, Florence Jaumotte, Longji Li, Giovanni Melina, Augustus Panton, Carlo Pizzinelli, Emma Rockall, and Marina Mendes Tavares. 2024. Gen-AI: Artificial Intelligence and

the Future of Work. IMF Staff Discussion Note No. SDN/2024/001. International Monetary Fund. https://doi.org/10.5089/9798400262548.006

CB Insights. 2018. Artificial Intelligence Trends To Watch In 2018. CB Insights. https://www.cbinsights.com/reports/CB-Insights_State-of-Artificial-Intelligence-2018.pdf

CB Insights. 2025a. "Who's Winning the AI Coding Race?" CB Insights. https://www.cbinsights.com/research/report/ai-coding-race/

CB Insights. 2025b. The AI agent market map. CB Insights. https://www.cbinsights.com/research/ai-agent-market-map-2025/

CB Insights. 2025c. "AI Agent Bible: The Ultimate Guide to Agent Disruption." CB Insights. https://www.cbinsights.com/research/report/ai-agent-bible/

CB Insights. 2025d. “The Complete List of Unicorn Companies.” CB Insights. https://www.cbinsights.com/research-unicorn-companies

CBO 1979. The Effects of the Tokyo Round of Multilateral Trade Negotiations on the U.S. Economy: An Updated View. U.S. Government Printing Office.

CEA. 1979. The 1979 Joint Economic Report. U.S. Government Printing Office. https://www.jec.senate.gov/reports/96th%20Congress/The%201979%20Joint%20Economic%20Report%20(930).pdf

CEA. 1996. Economic Report of the President. Washington, D.C.: U.S. Government Printing Office.

CEA. 2024. What Drives the U.S. Services Trade Surplus? Growth in Digitally Enabled Services Exports. White House. https://bidenwhitehouse.archives.gov/cea/written-materials/2024/06/10/what-drives-the-u-s-services-trade-surplus-growth-in-digitally-enabled-services-exports/

Center for AI Safety. 2023. “AI Extinction Statement Press Release.” https://safe.ai/work/press-release-ai-risk

CFPB. 2023. Consumer Financial Protection Circular 2023-03: Adverse action notification requirements and the proper use of the CFPB’s sample forms provided in Regulation B. https://files.consumerfinance.gov/f/documents/cfpb_adverse_action_notice_circular_2023-09.pdf

Challapally, Aditya, Chris Pease, Ramesh Raskar, and Pradyumna Chari. 2025. “The GenAI Divide: State of AI in Business 2025.” MIT Media Lab, NANDA Initiative. https://nanda.media.mit.edu/ai_report_2025.pdf

Chalmers, Stephanie. 2025. “Commonwealth Bank Backtracks on AI Job Cuts, Apologises for ‘Error’ as Call Volumes Rise.” ABC News (Australia). https://www.abc.net.au/news/2025-08-21/cba-backtracks-on-ai-job-cuts-as-chatbot-lifts-call-volumes/105679492

Chander, Anupam, Margot E. Kaminski, and William McGeveran. 2021. “Catalyzing Privacy Law.” Minnesota Law Review 105: 1733-1802.

Chamberlin, Edward. 1927. The Theory of Monopolistic Competition. Unpublished Manuscript.

Chandler, Alfred, Jr. (1977). The Visible Hand: The Managerial Revolution in American Business. Cambridge, MA: Harvard University Press Press.

Chang, Wei, Priscilla Owusu-Mensah, Jordan Everson, and Chelsea Richwine. 2025. “Hospital Trends in the Use, Evaluation, and Governance of Predictive AI, 2023–2024.” ASTP Data Brief no. 80. Office of the Assistant Secretary for Technology Policy/Office of the National Coordinator for Health Information Technology. HealthIT.gov.

https://www.healthit.gov/data/data-briefs/hospital-trends-use-evaluation-and-governance-predictive-ai-2023-2024

Chari, Varadarajan, and Hugo Hopenhayn. 1991. "Vintage human capital, growth, and the diffusion of new technology." Journal of political Economy 99(6): 1142-1165.

Chavez, Anthony. 2024. "A New Path for Privacy Sandbox on the Web." Google The Keyword Blog. https://privacysandbox.com/intl/en_us/news/privacy-sandbox-update/

Chen, Mark, et al. 2021. “Evaluating Large Language Models Trained on Code.” arXiv preprint arXiv:2107.03374. https://arxiv.org/abs/2107.03374

Chevron. 1984. Chevron U.S.A., Inc. v. Natural Resources Defense Council, Inc., 467 U.S. 837. U.S. Supreme Court, June 25. https://supreme.justia.com/cases/federal/us/467/837

Chia, Austin. 2024. Open Neural Network Exchange (ONNX) Explained. Splunk. https://www.splunk.com/en_us/blog/learn/open-neural-network-exchange-onnx.html

Chicago Board of Trade. 1918. Board of Trade of Chicago v. United States, 246 U.S. 231. https://supreme.justia.com/cases/federal/us/246/231/

Chien, Colleen, Jorge Contreras, Thomas Cotter, Brian Love, Christopher Seaman, and Norman Siebrasse. 2019. "Enhanced Damages, Litigation Cost Recovery, and Interest." In Patent Remedies and Complex Products: Toward a Global Consensus. Edited by C. Bradford Biddle, Jorge Contreras, Brian Love, and Norman Siebrasse. Cambridge: Cambridge University Press, 90-114.

Chin, Josh, and Liza Lin. 2022. Surveillance State: Inside China's Quest to Launch a New Era of Social Control. New York: St. Martin's Press.

Chin, Josh, and Raffaele Huang (2025). “The AI Cold War That Will Redefine Everything.” The Wall Street Journal. https://www.wsj.com/tech/ai/the-ai-cold-war-that-will-redefine-everything-4e1810b2?msockid=099ba79ac5be67760541b13ac47066c7

Chollet, François 2019. On the Measure of Intelligence. arXiv preprint *arXiv:1911.01547*. https://ms456000.wpcomstaging.com/wp-content/uploads/2020/02/2020-02-25-d0a1d182d0b0d182d18cd18f-d0bfd180d0be-d098d098.pdf

Christiano, Paul, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. “Deep Reinforcement Learning from Human Preferences.” arXiv preprint arXiv:1706.03741. https://arxiv.org/abs/1706.03741

City of Arlington. 2013. City of Arlington v. Federal Communications Commission. No. 11-1545. US Supreme Court, May 20. https://supreme.justia.com/cases/federal/us/569/290

CJEU. 2024a. Google LLC and Alphabet Inc. v European Commission (Google Shopping). Case C-48/22 P. Judgment, 10 September 2024. https://curia.europa.eu/juris/document/document.jsf?docid=289925&doclang=EN.

CJEU. 2024b. European Commission v Ireland and Apple Sales International (Apple State Aid). Case C-465/20 P. Judgment, 10 September 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62020CJ0465

Clark, Kim. 1985. The interaction of design hierarchies and market concepts in technological evolution. Research Policy 14(5): 235-251.

Clark, Kate. 2025. “Nvidia Licenses Groq’s AI Technology as Demand for Cutting-Edge Chips Grows.” The Wall Street Journal. https://www.wsj.com/articles/nvidia-licenses-groqs-ai-technology-as-demand-for-cutting-edge-chips-grows-87990cbe856818d5eddac44c7b1cdeb8

Clayton, Lewis, and Eric Stone. 2018. "Claim Drafting, Damages Apportionment in Multi-Component Product Cases." New York Law Journal 259(8). https://www.paulweiss.com/media/cvbjgfyj/11jan2018nylj.pdf

Clifford, Stephanie, and Quentin Hardy. 2013. "Attention, Shoppers: Store Is Tracking Your Cell." New York Times. https://www.nytimes.com/2013/07/15/business/attention-shopper-stores-are-tracking-your-cell.html

Clinton White House. 1995. National Economic Council Function. The White House. https://clintonwhitehouse5.archives.gov/WH/EOP/nec/html/function.html

Clinton, William J. 1993. Executive Order 12866: Regulatory Planning and Review. Federal Register 58(190): 51735-51744.

Clinton, William, and Albert Gore. 1997. A Framework for Global Electronic Commerce. The White House. https://clintonwhitehouse3.archives.gov/WH/New/Commerce/read.html

Clinton, William. 1993. Executive Order 12866—Regulatory Planning and Review. 58 FR: 51735. https://www.archives.gov/federal-register/executive-orders/pdf/12866.pdf

CMA. 2020. Online Platforms and Digital Advertising Market Study: Final Report. London: Competition and Markets Authority. https://assets.publishing.service.gov.uk/media/5efc57ed3a6f4023d242ed56/Final_report_1_July_2020_.pdf

Coase, Ronald. 1937. “The Nature of the Firm.” Economica, New Series 4(16): 386–405.

Coase, Ronald. 1960. "The Problem of Social Cost." Journal of Law and Economics 3 (October): 1-44.

Cobb, Charles, and Paul Douglas. 1928. "A Theory of Production." American Economic Review 18(1): 139-165.

Cochrane, Willard. 1979. The Development of American Agriculture. Minneapolis: University of Minnesota Press.

Cohen, Ben. 2025. “It’s Waymo’s World. We’re All Just Riding in It.” The Wall Street Journal. https://www.wsj.com/tech/waymo-cars-self-driving-robotaxi-tesla-uber-0777f570

Cohen, Benjamin and Daniel Sisle. 1970. Export of developing countries in the 1960's. Discussion Paper No. 94, Yale University, Economic Growth Center.

Cohen, Julie, and Mark Lemley. 2001. "Patent Scope and Innovation in the Software Industry." California Law Review 89(1): 1–57.

Cohen, Wesley, Richard Nelson, and John Walsh. 2002. "Links and Impacts: The Influence of Public Research on Industrial R&D." Management Science 48(1): 1-23.

Columbus, Louis 2017. McKinsey's State Of Machine Learning And AI, 2017. https://www.forbes.com/sites/louiscolumbus/2017/07/09/mckinseys-state-of-machine-learning-and-ai-2017/

Comai, Sam. 2025. “AI Pilot Programs in K-12 Settings.” Education Commission of the States. ECS. https://www.ecs.org/ai-artificial-intelligence-pilots-k12-schools/

Comin, Diego, and Bart Hobijn. 2010. "An Exploration of Technology Diffusion." American Economic Review 100(5): 2031-2059.

Compulife. 2020. Compulife Software Inc. v. Newman, 959 F.3d 1288 (11th Cir.). https://law.justia.com/cases/federal/appellate-courts/ca11/18-12004/18-12004-2020-05-20.html

Constine, Josh. 2018. "Facebook Cuts Off Data Brokers to Avoid Scandals Like Cambridge Analytica's." TechCrunch. https://techcrunch.com/2018/03/28/facebook-will-cut-off-access-to-third-party-data-for-ad-targeting

Continental Can. 1964. United States v. Continental Can Co., No. 367. US Supreme Court, June 22. https://supreme.justia.com/cases/federal/us/378/441

Continental T.V. 1977. Continental T.V., Inc. v. GTE Sylvania Inc., No. 76-15. US Supreme Court, June 23. https://supreme.justia.com/cases/federal/us/433/36

Contino. 2021. What is compliance as code? Benefits, use cases and tools. https://www.contino.io/insights/compliance-as-code

Contreras, Jorge. 2015. “A Brief History of FRAND: Analyzing Current Debates in Standard-Setting and Antitrust Law.” Antitrust Law Journal 80(1): 39–120.

Contreras, Jorge. 2017. "From Private Ordering to Public Law." Harvard Journal of Law and Technology 30(Special Symposium): 211-231.

Contreras, Jorge. 2022. “Technical Standards: Fair, Reasonable and Non-Discriminatory (FRAND) Licensing.” In Intellectual Property Licensing and Transactions, edited by Jorge L. Contreras. Cambridge, UK: Cambridge University Press: 196–247.

Cook, Lisa. 2024. “Artificial Intelligence, Big Data, and the Path Ahead for Productivity.” Speech, Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/newsevents/speech/cook20241001a.htm

Copeland, Curtis. 2013. Federal Rulemaking. CRS Report RL32397. https://www.congress.gov/crs-product/RL32397

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2009. “Intangible Capital and U.S. Economic Growth.” Review of Income and Wealth 55(3): 661–685.

CoSN. 2023. An Introduction to Interoperability Standards for Education Leaders: Strategically Connecting the K–12 Enterprise. Washington, DC: Consortium for School Networking. https://www.cosn.org/wp-content/uploads/2023/06/COSN_Interop_Intro_F3.pdf

Cowen, Tyler. 2011. The Great Stagnation. New York: Dutton.

Cowen, Tyler. 2025. “Why I Think AI Take-off Is Relatively Slow.” Marginal Revolution. https://marginalrevolution.com/marginalrevolution/2025/02/why-i-think-ai-take-off-is-relatively-slow.html

Cramton, Peter. 1997. The FCC Spectrum Auctions: An Early Assessment. Journal of Economics and Management Strategy 6(3): 431–495.

Cramton, Peter. 2010. "700 MHz Device Flexibility Promotes Competition." Working Paper. www.cramton.umd.edu/papers2010-2014/cramton-700-mhz-device-flexibility-promotes-competition.pdf

Crandall, Robert, and Clifford Winston. 2003. Does Antitrust Policy Improve Consumer Welfare? Assessing the Evidence. Journal of Economic Perspectives 17(4): 3-26.

Crandall, Robert. 2001. "Managed Competition in U.S. Telecommunications." The Brookings Review 19(2): 19–21.

Crane, Daniel. 2007. Antitrust Modesty. Michigan Law Review 105(8): 1193-1212.

Crawford, Susan. 2013. Captive Audience. New Haven, CT: Yale University Press.

CRFCAS. 1975. "Structure and Internal Procedures: Recommendations for Change." CRFCAS. https://www.fordlibrarymuseum.gov/library/document/0019/4520540.pdf

Crissinger, Cathy. 2000. "AT&T Corp. v. Excel Communications, Inc." Berkeley Tech. Law Journal 15(165): 165–184.

Crouzet, Nicolas, and Janice C. Eberly. 2019. “Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles.” NBER Working Paper No. 25869. Cambridge, MA: National Bureau of Economic Research.

Crémer, Jacques, Yves-Alexandre De Montjoye, and Heike Schweitzer. 2019. Competition policy for the digital era. Brussels: Publications Office of the European Union.

CSIRO. 2015. Commonwealth Scientific and Industrial Research Organisation v. Cisco Systems, Inc., No. 15-1066. US Court of Appeals for the Federal Circuit, December 1. https://cafc.uscourts.gov/sites/default/files/opinions-orders/15-1066.OPINION.12-1-2015.1.PDF

Cunningham, Waylon 2024. Big Mac goes Big Tech, with a few hiccups. Reuters. https://www.reuters.com/business/retail-consumer/big-mac-goes-big-tech-with-few-hiccups-2024-03-16/

Cuofano, Genaro. 2025. AI Companies Hit $2.8M Revenue Per Employee: The Death of Traditional Business Models. FourWeekMBA. https://fourweekmba.com/ai-companies-hit-2-8m-revenue-per-employee-the-death-of-traditional-business-models/

Custers, Bart, Henning Lahmann and Benjamyn Scott. 2025. From liability gaps to liability overlaps: Shared responsibilities and fiduciary duties in AI and other complex technologies. AI and Society 40(5): 4035–4050.

Cutter, Chip, and Haley Zimmerman. 2025. CEOs Start Saying the Quiet Part Out Loud: AI Will Wipe Out Jobs. The Wall Street Journal. https://www.wsj.com/tech/ai/ai-white-collar-job-loss-b9856259

Cyber Creative Institute. 2013. "Evaluation of LTE Essential Patents Declared to ETSI." Version 3.0. Cyber Creative Institute. https://www.cybersoken.com/wp_kanri/wp-content/uploads/2011/12/lte03EN.pdf

Cyberspace Administration of China. 2025. "关于印发《人工智能生成合成内容标识办法》的通知." https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

Dabbir, Rupesh. 2025. The future of code: How AI is transforming software development. Forbes. https://www.forbes.com/councils/forbestechcouncil/2025/04/04/the-future-of-code-how-ai-is-transforming-software-development/

Dadaboyev, Sherzodbek Murodilla Ugli, Jasmina Abdullayeva, Naval Abbosova, Afina Suleymenova and Komila Mamadjanova. 2025. “Role of artificial intelligence in employee recruitment: systematic review and future research directions.” Discover Global Society 3(99).

Daffodil Software. 2025. Top 10 Image Recognition APIs in 2025. https://insights.daffodilsw.com/blog/top-10-image-recognization-apis-for-app-development

Dahlman, Erik, Stefan Parkvall, and Johan Sköld. 2011. 4G: LTE/LTE-Advanced for Mobile Broadband. Oxford: Academic Press.

Dally, Bill. 2023. Heeding Huang’s Law: Video Shows How Engineers Keep the Speedups Coming. NVIDIA Blog. https://blogs.nvidia.com/blog/huangs-law-dally-hot-chips/

Damle, Sy, Alli Stillman, Britt Lovejoy, Ivana Dukanovic. 2024. "The ELVIS Act: Tennessee Shakes Up Its Right of Publicity Law and Takes On Generative AI." Latham and Watkins LLP. https://www.lw.com/admin/upload/SiteAttachments/The-ELVIS-Act-Tennessee-Shakes-Up-Its-Right-of-Publicity-Law-and-Takes-On-Generative-AI.pdf

Dance, Gabriel, Michael LaForgia, and Nicholas Confessore. 2018. "As Facebook Raised a Privacy Wall, It Carved an Opening for Tech Giants." The New York Times. https://www.nytimes.com/2018/12/18/technology/facebook-privacy-data-sharing-ebay-microsoft-amazon-netflix-spotify.html

Darveau-Garneau, Matt, Manuel Frey, Beth Goldberg, and Nick South. 2022. "As the Cookie Crumbles: Three Strategies for Advertisers to Thrive." McKinsey & Company.

https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/as-the-cookie-crumbles-three-strategies-for-advertisers-to-thrive

Dastin, Jeffrey. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Databricks. 2024. Databricks Unified Data and AI Platform. https://databricks.com/

DataCenterDynamics. 2025. Iron Mountain breaks ground on first data center in Miami. DataCenterDynamics. https://www.datacenterdynamics.com/en/news/iron-mountain-breaks-ground-on-first-data-center-in-miami/

David, Paul. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." American Economic Review 80(2): 355-361.

Davies, Paul. 2023. “Potentially Disastrous for Innovation: Tech Sector Reacts to the EU AI Act.” Euronews. https://www.euronews.com/next/2023/12/15/potentially-disastrous-for-innovation-tech-sector-says-eu-ai-act-goes-too-far

Daws, Ryan. 2025. "AI and its impact on software development jobs." Developer Tech. https://www.developer-tech.com/news/ai-impact-on-software-development-jobs/

de Boer. 2025. Nvidia Crushes Competition With 94% GPU Market Share. Yahoo Finance. https://finance.yahoo.com/news/nvidia-crushes-competition-94-gpu-144404532.html

De Loecker, Jan, and Jan Eeckhout. 2017. "The Rise of Market Power and the Macroeconomic Implications." NBER Working Paper No. 23687. Cambridge, MA: National Bureau of Economic Research.

De Loecker, Jan, József Eeckhout, and Gabriel Unger. 2020. “The Rise of Market Power and the Macroeconomic Implications.” Quarterly Journal of Economics 135(2): 561–644.

Dean, Jeffrey, et al. 2012. "Large Scale Distributed Deep Networks." Advances in Neural Information Processing Systems 25 (NIPS 2012): 1223–1231.

Dedrick, Jason, Kenneth Kraemer and Greg Linden. 2010. "Who profits from innovation in global value chains?: a study of the iPod and notebook PC." Industrial and Corporate Change 19(1): 81-116.

DeLamarter, Richard. 1986. Big Blue. New York: Dodd, Mead and Company.

Delrahim, Makan 2018. DoJ Business Review Letter for Avanci LLC. DoJ. https://www.justice.gov/atr/page/file/1298626/download

Demsetz, Harold 1973. Industry Structure, Market Rivalry, and Public Policy. The Journal of Law and Economics 16(1): 1-9.

Dentsply International. 2005. United States v. Dentsply International, Inc., No. 03-4097, 3d Cir., Feb. 24. https://law.justia.com/cases/federal/appellate-courts/F3/399/181/527982

Desmarais, Anna. 2025. “Two city governments in Denmark are moving away from Microsoft amid Trump and US Big Tech concerns.” Euronews Next. https://www.euronews.com/next/2025/06/12/two-city-governments-in-denmark-are-moving-away-from-microsoft-amid-trump-and-us-big-tech-

Destler, I. M. (Mac) 1996. The National Economic Council. Washington, DC: Peterson Institute for International Economics.

Deuel. 1995. In re Deuel. No. 94-1202. US Federal Circuit Court, March 28. https://law.justia.com/cases/federal/appellate-courts/F3/51/1552/617820/

Deutsche Telekom 2020. New York v. Deutsche Telekom AG, No. 1:19-cv-05434, 2020 WL 1879475 (S.D.N.Y.). https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2019cv05434/517350/409

Devine, Warren. 1983: "From Shafts to Wires: Historical Perspectives on Electrification." The Journal of Economic History 43(2): 347-372.

Dezan Shira and Associates. 2024. “Pre-Tax Super Deduction of R&D Expenses in China – An Explainer.” China Briefing. https://www.china-briefing.com/news/china-rd-expenses-pre-tax-super-deduction-explainer/

Diab, Wael William, and Mike Mullane. 2024. “How the ISO and IEC Are Developing International Standards for the Responsible Adoption of AI.” UNESCO. https://www.unesco.org/en/articles/how-iso-and-iec-are-developing-international-standards-responsible-adoption-ai

Diamond. 1981. Diamond, Commissioner of Patents and Trademarks v. Diehr et al. No. 79-1112. US Supreme Court, March 3. https://supreme.justia.com/cases/federal/us/450/175/

Dickinson, Grace. 2023. Are AI-powered robots the future of dishwashing in the restaurant industry? Back of House. https://backofhouse.io/resources/ai-powered-robots-future-dishwashing-restaurant-industry-automation-technology

DICOM Standards Committee. 2025. “Artificial Intelligence and DICOM (WG-23): Interoperability Guidance and Resources.” DICOM Standard. https://www.dicomstandard.org/ai/

Disu, Opemipo. 2024. Amazon Bedrock. Microtica (Dev.to). https://dev.to/microtica/amazon-bedrock-a-practical-guide-for-developers-and-devops-engineers-4k8g

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. “Documenting Large Webtext Corpora.” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 1286–1305. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.98.

DoE. 2015. Frequently Asked Questions—Cloud Computing. Privacy Technical Assistance Center (PTAC). DoE. https://studentprivacy.ed.gov/sites/default/files/resource_document/file/FAQ_Cloud_Computing_0.pdf

Doe. 2016. Doe No. 1 v. Backpage.com, LLC. No. 15-1724. US Court of Appeals, First Circuit, March 14. https://law.justia.com/cases/federal/appellate-courts/ca1/15-1724/15-1724-2016-03-14.html

DoJ and FTC. 1992. Horizontal Merger Guidelines. https://www.justice.gov/archives/atr/1992-merger-guidelines

DoJ and FTC. 1995. Antitrust Guidelines for the Licensing of Intellectual Property. DoJ and FTC. https://www.justice.gov/atr/archived-1995-antitrust-guidelines-licensing-intellectual-property

DoJ and FTC. 2010. "Horizontal Merger Guidelines." DoJ and FTC. https://www.justice.gov/atr/horizontal-merger-guidelines-08192010

DOJ and USPTO. 2025. "Statement of Interest of the United States." Radian Memory Systems LLC v. Samsung Electronics Co., No. 2:24-cv-1073 (E.D. Tex.). https://justice.gov/atr/media/1404506/dl

DoJ. 1982. Merger Guidelines. https://www.justice.gov/archives/atr/1982-merger-guidelines

DoJ. 1984. "Non-Horizontal Merger Guidelines." https://www.justice.gov/archives/atr/1984-merger-guidelines

DoJ. 1988. Antitrust Enforcement Guidelines for International Operations. Washington, D.C.: DoJ.

DoJ. 1997. "Business Review Letter to MPEG LA." DoJ Antitrust Division. Washington, D.C.: DoJ.

DoJ. 2004. "Justice Department Implements the Standards Development Organization Advancement Act of 2004." Press Release. https://www.justice.gov/archive/atr/public/press_releases/2004/204345.htm

DoJ. 2008. Competition and Monopoly: Single-Firm Conduct Under Section 2 of the Sherman Act. Washington, DC: DoJ.

DoJ. 2009. Justice Department Withdraws Report on Antitrust Monopoly Law. Press Release. https://www.justice.gov/archives/opa/pr/justice-department-withdraws-report-antitrust-monopoly-law

DoJ. 2015. "Business Review Letter to Institute of Electrical and Electronics Engineers, Incorporated." Press Release. https://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/02/02/ieee_business_review_letter.pdf

DoJ. 2019. “Complaint.” United States v. Sabre Corp., et al., No. 1:19-cv-01548 (D. Del.), filed Aug. 20, 2019. Washington, D.C.: DoJ.

DoJ. 2020. “Plaintiff’s Pretrial Brief [Redacted Public Version].” United States v. Sabre Corp., et al., No. 1:19-cv-01548 (D. Del.), filed Jan. 15, 2020. Washington, D.C.: DoJ.

DoJ. n.d.a. Criminal Enforcement. https://www.justice.gov/atr/criminal-enforcement

DoJ. n.d.b. Civil Enforcement. https://www.justice.gov/atr/civil-enforcement

DoJ/USPTO 2025. "Joint Comment on the Public Interest in Certain DRAM Devices, Products Containing the Same and Components Thereof." ITC Investigation No. 337-TA-3854. https://justice.gov/atr/media/1419496/dl

DoJ/USPTO/NIST. 2019. Policy Statement on Remedies for Standards-Essential Patents Subject to Voluntary F/RAND Commitments. DoJ/USPTO/NIST https://www.justice.gov/atr/page/file/1228016/dl?inline

DoJ/USPTO/NIST. 2022. Withdrawal of 2019 Policy Statement on Remedies for Standards-Essential Patents Subject to Voluntary F/RAND Commitments. DoJ/USPTO/NIST. https://www.uspto.gov/sites/default/files/documents/SEP2019-Withdrawal.pdf

DoL. 2024. History of Federal Minimum Wage Rates Under the Fair Labor Standards Act, 1938–2009. https://www.dol.gov/agencies/whd/minimum-wage/history/chart

Domen. 2021. Domen v. Vimeo, Inc. No. 20-616. US Court of Appeals, Second Circuit, March 11. https://law.justia.com/cases/federal/appellate-courts/ca2/20-616/20-616-2021-03-11.html

Domingos, Pedro. 2015. The Master Algorithm. New York: Basic Books.

Donnellan, Douglas, Andy Lawrence, Daniel Bizo, Peter Judge, John O'Brien, Jacqueline Davis, Max Smolaks, Jabari Williams-George, and Rose Weinschenk. 2024. Uptime Institute Global Data Center Survey 2024. Uptime Institute. https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/2024.GlobalDataCenterSurvey.Report.pdf

DoS. 1980. 1980 Editorial Note: Executive Order 12188 and the Trade Reorganization Plan. Foreign Relations of the United States, 1977–1980, Volume XXVIII. Organization and Management of Foreign Policy. Document 121. https://history.state.gov/historicaldocuments/frus1977-80v28/d121

Doss, Christopher, Robert Bozick, Heather Schwartz, Lisa Chu, Lydia Rainey, Ashley Woo, Justin Reich, Jesse Dukes. 2025. AI Use in Schools Is Quickly Increasing but Guidance Lags Behind: Findings from the RAND Survey Panels. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RRA4180-1.html

Dr. Miles Medical. 1911. Dr. Miles Medical Co. v. John D. Park and Sons Co., 220 U.S. 373. https://supreme.justia.com/cases/federal/us/220/373

Draghi, Mario. 2024. The Future of European Competitiveness (Executive Summary Highlights). European Commission. https://commission.europa.eu/document/download/97e481fd-2dc3-412d-be4c-f152a8232961_en

Drew, Rocket. 2025. AI Agenda. The Information. https://www.theinformation.com/articles/chinese-tech-behind-amazons-humanoid-robots?rc=tuorsb

Dreyfuss, Rochelle Cooper. 1989. "The Federal Circuit: A Case Study in Specialized Courts." New York University Law Review 64(1): 1–77.

Dudley, Susan. 2020. "OMB's Circular A-4 and the History of Regulatory Impact Analysis." Administrative Law Review 72(1): 11–32.

Duffer, Sara, Singh Ashish, Denis Batalov, and Peter Hallinan. 2025. "Building trust in AI: The AWS approach to the EU AI Act." AWS Machine Learning Blog. https://aws.amazon.com/blogs/machine-learning/building-trust-in-ai-the-aws-approach-to-the-eu-ai-act

Dumont, Andrew. 2024. “Changes in the U.S. Economy and Rural-Urban Employment Disparities.” FEDS Notes, Board of Governors of the Federal Reserve System (January 19, 2024). https://www.federalreserve.gov/econres/notes/feds-notes/changes-in-the-us-economy-and-rural-urban-employment-disparities-20240119.html [harvardmagazine.com]

Dupont. 1956. United States v. E. I. du Pont de Nemours and Co., 351 U.S. 377. https://supreme.justia.com/cases/federal/us/351/377

Durie, Daralyn, and Mark Lemley. 2008. "A Realistic Approach to the Obviousness of Inventions." William and Mary Law Review 50(3): 989–1019.

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." Foundations and Trends in Theoretical Computer Science 9(3–4): 211–407.

Dyroff. 2019. Dyroff v. Ultimate Software Group, Inc. No. 18-15175. US Court of Appeals, Ninth Circuit, August 20. https://law.justia.com/cases/federal/appellate-courts/ca9/18-15175/18-15175-2019-08-20.html

Dzieza, Josh, and Hayden Field. 2025. “Feeding the Machine.” The Verge. https://www.theverge.com/2025/12/15/feeding-the-machine-ai-training-data-mercor-scale-surge

D’Silva, Charmaine. 2020. “11 Weeks of Android: Privacy and Security.” Android Developers Blog. https://android-developers.googleblog.com/2020/06/11-weeks-of-android-privacy-and-security.html

Easterbrook, Frank. 1984. The Limits of Antitrust. Texas Law Review 63(1): 1-40.

Eastman Kodak. 1992. Eastman Kodak Co. v. Image Technical Services, Inc., 504 U.S. 451. https://supreme.justia.com/cases/federal/us/504/451

eBay. 2000. eBay, Inc. v. Bidder's Edge, Inc. No. C-99-21200 RMW. US District Court, N.D. California, May 24. https://law.justia.com/cases/federal/district-courts/FSupp2/100/1058/2478126

eBay. 2006. eBay, Inc. v. MercExchange, L.L.C., No. 05-130. US Supreme Court, May 15. https://supreme.justia.com/cases/federal/us/547/388

Edelman, Benjamin. 2015. "Does Google Leverage Market Power Through Tying and Bundling?" Journal of Competition Law and Economics 11(2): 365–400.

Edlin, Aaron. 2002. "Stopping Above-Cost Predatory Pricing." Yale Law Journal 111(4): 941–991.

EDM Council. n.d. Cloud Data Management Capabilities Framework. EDM Council. https://edmcouncil.org/

Edwards, Chris. 2005. Agricultural Policy. Cato Handbook on Policy (6th ed). https://www.cato.org/sites/cato.org/files/serials/files/cato-handbook-policymakers/2005/9/hb109-31.pdf

Efrati, Amir, and Stephanie Palazzolo. 2025. “OpenAI Targets AGI with System That Thinks Like a Pro Engineer.” The Information. https://www.theinformation.com/articles/openai-targets-agi-with-system-that-thinks-like-a-pro-engineer

EIA. 2018. “PURPA-qualifying capacity increases, but it’s still a small portion of added renewables.” Today in Energy. https://www.eia.gov/todayinenergy/detail.php?id=36912

Eichengreen, Barry. 1992. Golden Fetters: The Gold Standard and the Great Depression, 1919–1939. New York, NY: Oxford University Press.

Eisenach, Jeffrey and Robert Kulick. 2020. Economic impacts of mobile broadband innovation: Evidence from the transition to 4G. AEI Economics Working Paper 2020-05. American Enterprise Institute. https://aei.org/wp-content/uploads/2020/06/Eisenach-Kulick-Mobile-Broadband-Innovation-WP.pdf

Elhauge, Einer 2003. Defining Better Monopolization Standards. Stanford Law Review 56: 253-234.

Elmore, John. 2024. “The Foxconn Connection: Unraveling Apple’s Outsourcing Strategy.” TheTechyLife. https://thetechylife.com/why-does-apple-outsource-to-foxconn/

Elsner, Mark, Grace Atkinson, and Saadia Zahidi. 2025. The Global Risks Report 2025: 20th Edition. World Economic Forum. https://www.weforum.org/publications/global-risks-report-2025/

Elzinga, Kenneth, and Thomas Hogarty. 1973. "The Problem of Geographic Market Delineation in Antimerger Suits." The Antitrust Bulletin 18(1): 45–81.

Emberson, Luke. 2025. Frontier open models may surpass 10^{26} FLOP of training compute before 2026. Epoch AI. https://epochai.org/data-insights/frontier-open-models-may-surpass-10-26-flop-of-training-compute-before-2026

Emory Solutions. 2024. The Problem with AI Hallucinations: Why They Happen, How to Prevent, and What Risks Exist. Emory Solutions. https://www.emorysolutions.com/the-problem-with-ai-hallucinations-why-they-happen-how-to-prevent-and-what-risks-exist/

Engler, Alex 2023. "The Evolution of AI Regulation in the United States." Washington, DC: Brookings Institution Report.

EPA. 2014. Regulatory Impact Analysis: Control of Air Pollution from Motor Vehicles: Tier 3 Motor Vehicle Emission and Fuel Standards. EPA-420-R-14-005. https://www.govinfo.gov/content/pkg/FR-2014-04-28/pdf/2014-06954.pdf

Epoch AI. 2024. "Data on Machine Learning Hardware." Epoch AI. https://epoch.ai/data/machine-learning-hardware

Epstein, Richard. 2006. How Progressives Rewrote the Constitution. Washington, D.C.: Cato Institute.

ETSI. 2024. ETSI IPR Policy. Sophia Antipolis: ETSI.

ETSI. 2019. "Intellectual Property Rights Policy." ETSI Rules of Procedure, Annex 6. Sophia Antipolis: European Telecommunications Standards Institute. https://portal.etsi.org/directives/40_directives_apr_2019.pdf

EU General Court. 2022. Google and Alphabet v. European Commission. Case T-604/18, ECLI:EU:T:2022:541. September 14 2022. https://curia.europa.eu/juris/document/document.jsf?docid=265435

EU General Court. 2024. Google and Alphabet v European Commission (Google AdSense for Search). Case T-334/19. Judgment, 18 September 2024. https://eur-lex.europa.eu/eli/C/2024/6420/oj/eng

EU. 2022. Regulation (EU) 2022/1925 (Digital Markets Act). OJ L 265, 12.10.2022, 1–66. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R1925

EU. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

EuroHPC JU. 2024. Consolidated Annual Activity Report 2024. https://www.eurohpc-ju.europa.eu/document/download/a8889156-6b46-4e91-bf70-e4baf5cdb68f_en

European Commission. 2024. "Commission sends Statement of Objections to Microsoft over possibly abusive tying practices regarding Teams." Press release. https://ec.europa.eu/commission/presscorner/api/files/document/print/en/ip_24_3446/IP_24_3446_EN.pdf

European Commission. 2025a. “Commission launches two strategies to speed up AI uptake in European industry and science.” Press release. https://ec.europa.eu/commission/presscorner/api/files/document/print/en/ip_25_2299/IP_25_2299_EN.pdf

European Commission. 2025b. "Commission finds Apple and Meta in breach of the Digital Markets Act." Press release. https://ec.europa.eu/commission/presscorner/api/files/document/print/en/ip_25_1085/IP_25_1085_EN.pdf

European Court of Auditors. 2025. Special Report 12/2025 — The EU's strategy for microchips: Ambitious targets, but funding and implementation risks remain. Luxembourg: Publications Office of the European Union. https://www.eca.europa.eu/ECAPublications/SR-2025-12/SR-2025-12_EN.pdf

European Parliament. "Artificial Intelligence Act." European Parliament legislative resolution of 13 March 2024, P9_TA(2024)0138. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=EP%3AP9_TA%282024%290138

European Union. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on Artificial Intelligence (AI Act). Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AL_202401689

Evans, Benedict. 2018. "Presentation: The End of the Beginning." Annual presentation. a16z Summit. https://www.ben-evans.com/benedictevans/2018/11/16/the-end-of-the-beginning

Evans, David. 2008. Antitrust Issues Raised by the Emerging Global Internet Economy. Northwestern University Law Review 102(4): 1987-2008.

Evans, David, and Richard Schmalensee. 2001. "Some Economic Aspects of Antitrust Analysis in Dynamically Competitive Industries." NBER Working Paper No. 8268. Cambridge, MA: National Bureau of Economic Research.

Evans, David, and Richard Schmalensee. 2016. Matchmakers. Boston, MA: Harvard Business Review Press.

Evans, David, Andrei Hagiu, and Richard Schmalensee. 2006. Invisible Engines. Cambridge, MA: MIT Press.

Exmark Manufacturing. 2018. Exmark Manufacturing Co. v. Briggs and Stratton Power Products Group, LLC. No. 2016-2197. US Court of Appeals, Federal Circuit, January 12. https://law.justia.com/cases/federal/appellate-courts/cafc/16-2197/16-2197-2018-01-12.html

Facebook 2022. "About Lookalike Audiences." Meta for Business. https://www.facebook.com/business/help/164749007013531

Facebook, Inc. 2021. “Annual Report 2020 (Form 10-K).” https://www.sec.gov/Archives/edgar/data/1326801/000132680121000014/fb-20201231.htm

Facebook. 2007a. “Facebook Unveils Platform for Developers of Social Applications.” Facebook Newsroom. https://about.fb.com/news/2007/05/facebook-unveils-platform-for-developers-of-social-applications/

Facebook. 2007b. “Facebook Develops Network Portals, New Inbox and Updates Site Design.” Facebook Newsroom. https://about.fb.com/news/2007/04/facebook-develops-network-portals-new-inbox-and-updates-site-design/

Facebook. 2007c. “Facebook Unveils Facebook Ads.” Facebook Newsroom. https://about.fb.com/news/2007/11/facebook-unveils-facebook-ads/

Facebook. 2013. "Announcing News Feed FYI: A Series of Blogs on News Feed Ranking." Facebook Newsroom. https://about.fb.com/news/2013/08/announcing-news-feed-fyi-a-series-of-blogs-on-news-feed-ranking/

Facebook. 2015. “Meta Blueprint (Facebook Blueprint): Learn New Skills to Build Your Brand or Business.” Meta Blueprint (training portal). https://www.facebookblueprint.com/student/catalog

Facebook. 2017. “Celebrating 5 Million Active Advertisers.” Facebook Newsroom. https://about.fb.com/news/2017/04/celebrating-5-million-active-advertisers

Facebook. 2019. In re Facebook, Inc., Consumer Privacy User Profile Litigation. No. 18-md-02843-VC. US District Court, N.D. California, September 9. https://law.justia.com/cases/federal/district-courts/california/candce/5:2018md02843/376023/233

Facebook. 2025. "About Custom Audiences." Meta Business Help Center. https://www.facebook.com/business/help/744354708981227

Fair Housing Council. 2008. Fair Housing Council of San Fernando Valley v. Roommates.com, LLC. No. 04-57115. US Court of Appeals, Ninth Circuit, April 3. https://law.justia.com/cases/federal/appellate-courts/F3/521/1157

Fairfield, Joshua. 2017. Owned. New York: Cambridge University Press.

Fairless, Tom, and David Luhnow. 2025. “The Tech Industry Is Huge—and Europe’s Share of It Is Very Small.” Wall Street Journal. https://www.wsj.com/tech/europe-big-tech-ai-1f3f862c

Farrell, Joseph, and Carl Shapiro. 2010. "Antitrust Evaluation of Horizontal Mergers." B.E. Journal of Theoretical Economics 10(1): Article 9.

Farrell, Joseph, and Garth Saloner. 1985. "Standardization, Compatibility, and Innovation." The RAND Journal of Economics 16(1): 70-83.

Farrell, Joseph, and Paul Klemperer. 2007. "Coordination and Lock-In: Competition with Switching Costs and Network Effects." In Handbook of Industrial Organization, Vol. 3, edited by Mark Armstrong and Robert Porter. Amsterdam: Elsevier, 1967-2072.

Farrell, Joseph, and Timothy Simcoe. 2012. "Choosing the Rules for Formal Standardization." Review of Industrial Organization 40(3): 159–88.

FCC. 2003. Review of the Section 251 Unbundling Obligations of Incumbent Local Exchange Carriers. Report and Order and Order on Remand and Further Notice of Proposed Rulemaking (FCC 03-36). https://docs.fcc.gov/public/attachments/FCC-03-36A1.pdf

FCC. 2005. Appropriate Framework for Broadband Access to the Internet over Wireline Facilities. Report and Order and Notice of Proposed Rulemaking (FCC 05-150). https://docs.fcc.gov/public/attachments/FCC-05-150A2.pdf

FCC. 2008. Auction 73: 700 MHz Band. Web page. https://www.fcc.gov/auction/73

FCC. 2009. Petition for Declaratory Ruling to Clarify Provisions of Section 332(c)(7)(B) to Ensure Timely Siting Review. Declaratory Ruling (FCC 09-99). https://apps.fcc.gov/edocs_public/attachmatch/FCC-09-99A1.pdf

FCC. 2013. Annual Assessment of the Status of Competition in the Market for the Delivery of Video Programming. Report (FCC 13-99). https://docs.fcc.gov/public/attachments/FCC-13-99A1.pdf

FCC. 2014. Protecting and Promoting the Open Internet. Notice of Proposed Rulemaking (FCC 14-61). https://apps.fcc.gov/edocs_public/attachmatch/FCC-14-61A1_Rcd.pdf

Federal Reserve Bank of St. Louis. 2024. Private Business Sector: Total Factor Productivity. Retrieved January 26, 2025, from https://fred.stlouisfed.org/series/MPU4900013

Federico, Giovanni, and Antonio Tena-Junguito. 2018. "World Trade, 1800-1938: A New Synthesis." Revista de Historia Económica / Journal of Iberian and Latin American Economic History 37(1): 9-41.

Federighi, Craig. 2019. “Apple previews iOS 13.” Apple Newsroom. https://www.apple.com/newsroom/2019/06/apple-previews-ios-13/ [apple.com]

Feigenbaum, James, and Daniel Gross. 2022. "Answering the Call of Automation: How the Labor Market Adjusted to the Mechanization of Telephone Operation." NBER Working Paper No. 28061. Cambridge, MA: National Bureau of Economic Research.

Feist. 1991. Feist Publications, Inc. v. Rural Telephone Service Co. No. 89-1909. US Supreme Court, March 18. https://supreme.justia.com/cases/federal/us/499/340/

Feldman. 2007. Feldman v. Google, Inc. No. 06-cv-02570-JG. US District Court, E.D. Pennsylvania, March 30. https://law.justia.com/cases/federal/district-courts/pennsylvania/paedce/2:2006cv02570/206160/46

FHWA (Federal Highway Administration). 2002. Appendix A Economic Effects of Transportation: The Freight Story. https://ops.fhwa.dot.gov/freight/freight_analysis/improve_econ/appa.htm

Field, Hayden. 2025a. "Google agrees to new $1 billion investment in Anthropic." CNBC. https://www.cnbc.com/2025/01/22/google-agrees-to-new-1-billion-investment-in-anthropic.html

Field, Hayden. 2025b. “Perplexity AI wrapping $500 million raise at $14 billion valuation.” CNBC. https://www.cnbc.com/2025/05/12/perplexity-funding-round-comet.html

Fielding, Sarah. 2023. Amazon says its new AI-powered robots reduce fulfilment time by 25 percent. Engadget. https://www.engadget.com/amazon-says-its-new-ai-powered-robots-reduce-fulfilment-time-by-25-percent-122517342.html

Filippucci, Francesco, Peter Gal, and Cecilia Susanna Jona Lasinio. 2024. "The impact of Artificial Intelligence on productivity, distribution and growth." Working Paper. https://iris.luiss.it/handle/11385/240418

Filistrucchi, Lapo, Damien Geradin, Eric Van Damme, and Pauline Affeldt. "Market definition in two-sided markets: Theory and practice." Journal of Competition Law and Economics 10(2): 293-339.

FillApp. 2025. The State of AI Browser Agents in 2025. FillApp Blog. https://fillapp.ai/blog/the-state-of-ai-browser-agents-2025/

Findlay, Ronald, and Kevin O'Rourke. 2007. Power and Plenty: Trade, War, and the World Economy in the Second Millennium. Princeton: Princeton University Press.

Finn, Chelsea, Sergey Levine, and Pieter Abbeel. "Guided cost learning: Deep inverse optimal control via policy optimization." In International conference on machine learning (PMLR, 2016): 49-58.

FinRegLab. 2025. Advancing the credit ecosystem: Machine learning and cash flow data in consumer underwriting. Empirical White Paper. FinRegLab. https://finreglab.org/research/advancing-the-credit-ecosystem-machine-learning-cash-flow-data-in-consumer-underwriting/

Fishback, Price. 2007. “The New Deal.” The Concise Encyclopedia of Economics. Indianapolis: Liberty Fund.

Fishman, Charles. 2006. The Wal-Mart Effect. New York: Penguin.

Fitch, Asa. 2025. “AI’s Big Leaps Are Slowing. That Could Be a Good Thing.” Wall Street Journal. https://www.wsj.com/tech/ai/ais-big-leaps-are-slowing-that-could-be-a-good-thing-34c87619

Force. 2019. Force v. Facebook, Inc. No. 18-397. US Court of Appeals, Second Circuit, July 31. https://law.justia.com/cases/federal/appellate-courts/ca2/18-397/18-397-2019-07-31.html

Foremost Pro Color. 1983. Foremost Pro Color, Inc. v. Eastman Kodak Co., 703 F.2d 534. https://law.justia.com/cases/federal/appellate-courts/F2/703/534/11835

Forgash, Emily, and Agnee Ghosh. 2025. "OpenAI, Nvidia Fuel $1 Trillion AI Market With Web of Circular Deals." Bloomberg News. https://www.bloomberg.com/news/features/2025-10-07/openai-s-nvidia-amd-deals-boost-1-trillion-ai-boom-with-circular-deals

Forlini, Emily. 2023. Samsung Software Engineers Busted for Pasting Proprietary Code Into ChatGPT. PC Magazine. https://www.pcmag.com/news/samsung-software-engineers-busted-for-pasting-proprietary-code-into-chatgpt

Free Enterprise Fund 2010. Free Enterprise Fund and Beckstead and Watts, LLP v. Public Company Accounting Oversight Board et al., No. 08-861, U.S. Supreme Court, June 28. https://scholar.google.com/scholar_case?case=6019616206791654633

Freeman, Mike. 2018. “FTC Antitrust Case Against Qualcomm Puts Business Model Under Microscope.” San Diego Union-Tribune. https://www.sandiegouniontribune.com/2018/12/30/ftc-antitrust-case-against-qualcomm-puts-business-model-under-microscope/

Freeman. 1978. In re Freeman. No. 77-598. US Court of Customs and Patent Appeals, March 30. https://digital-law-online.info/cases/197PQ464.htm

Frey, Carl Benedikt, and Michael Osborne. 2013. The Future of Employment: How Susceptible Are Jobs to Computerisation? Oxford Martin School. https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment

Frieden, Jeffrey. 2006. Global Capitalism: Its Fall and Rise in the Twentieth Century. New York: W. W. Norton.

Friedman, Milton, and Anna Jacobson Schwartz. 1963. A Monetary History of the United States, 1867–1960. Princeton, NJ: Princeton University Press.

Froeb, Luke. 2003. The Role of Expert Economic Testimony in Antitrust Litigation. FTC, Bureau of Economics. https://www.ftc.gov/sites/default/files/documents/public_statements/role-expert-economic-testimony-antitrust-litigation/031105froeb.pdf

FTC. 1966. In the Matter of National Dairy Products Corp. Order and Opinions. 70 F.T.C. 79. Docket No. 7018. https://www.ftc.gov/sites/default/files/documents/commission_decision_volumes/volume-70/ftcd-vol70july-december1966pages79-223.pdf

FTC. 1975. In the Matter of Xerox Corp. Decision and Order. 86 F.T.C. 364. Washington, DC: FTC.

FTC. 1982. Statement of the Federal Trade Commission Concerning Horizontal Mergers. Washington, DC: FTC.

FTC. 1996. In the Matter of Dell Computer Corp. Consent Order. 121 F.T.C. 616. Docket No. C-3658. https://www.ftc.gov/system/files/documents/cases/960617dellconsentorder.pdf

FTC. 2000. "Privacy of Consumer Financial Information; Final Rule." F.R. 65 (101): 33646–33689. https://www.govinfo.gov/content/pkg/FR-2000-05-24/html/00-12755.htm

FTC. 2002. In the Matter of Rambus Inc. Administrative Complaint. Docket No. 9302. FTC Matter/File No. 011 0017. https://www.ftc.gov/legal-library/browse/cases-proceedings/011-0017-rambus-inc-matter

FTC. 2003. To Promote Innovation: The Proper Balance of Competition and Patent Law and Policy. Washington, DC: FTC.

FTC. 2007. "Statement of Federal Trade Commission Concerning Google/DoubleClick." FTC File No. 071-0170. Washington, D.C.: FTC.

FTC. 2009. “Self-Regulatory Principles for Online Behavioral Advertising.” FTC Staff Report. https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-staff-report-self-regulatory-principles-online-behavioral-advertising/p085400behavadreport.pdf

FTC. 2012. In the Matter of Robert Bosch GmbH. Analysis to Aid Public Comment. Docket No. C-4377. FTC File No. 121-0081. https://www.ftc.gov/sites/default/files/documents/cases/2013/04/121126boschanalysis.pdf

FTC. 2013a. In the Matter of Motorola Mobility LLC and Google Inc. Decision and Order. Docket No. C-4410. FTC File No. 121-0120. https://www.ftc.gov/sites/default/files/documents/cases/2013/01/130103googlemotorolado.pdf

FTC. 2013b. In the Matter of Google Inc. Statement of the Federal Trade Commission Regarding Google’s Search Practices. Matter No. 111-0163. FTC File No. 111-0163. https://www.ftc.gov/legal-library/browse/cases-proceedings/public-statements/statement-federal-trade-commission-regarding-googles-search-practices-matter-google-inc

FTC. 2017a. "Statement of Federal Trade Commission's Acting Director of the Bureau of Competition on the Agency's Review of Amazon.com, Inc.'s Acquisition of Whole Foods Market Inc." Washington, D.C.: FTC.

FTC. 2017b. Cross-Device Tracking: A Federal Trade Commission Staff Report. https://www.ftc.gov/reports/cross-device-tracking-federal-trade-commission-staff-report-january-2017

FTC. 2025. A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority. https://www.ftc.gov/about-ftc/mission/enforcement-authority

FTC. n.d.a. The Antitrust Laws (Guide to Antitrust Laws). https://www.ftc.gov/advice-guidance/competition-guidance/guide-antitrust-laws/antitrust-laws

FTC. n.d.b. Clayton Act (Legal Library: Statutes). https://www.ftc.gov/legal-library/browse/statutes/clayton-act

FTSE Russell. 2025. Russell 3000 Index constituent list [Data set]. https://www.ftserussell.com/index-series/index-tools/russell-3000-index

Fu, Hui. 2025. Open-source LLMs surpass closed models in LMArena ranking. LinkedIn Post. https://www.linkedin.com/posts/huifu_the-latest-lmarena-ranking-of-open-source-activity-7383889353653035008-JsBJ

Fues, Eric. 2007. Implications of eBay v. MercExchange. Patent World. https://www.finnegan.com/en/insights/articles/implications-of-ebay-v-mercexchange.html

Funk, Jeffrey. 2009. “The Emerging Value Network in the Mobile Phone Industry: The Case of Japan and Its Implications for the Rest of the World.” Telecommunications Policy 33(1–2): 4–18.

Furman, Jason, and Peter Orszag. 2018. “A Firm-Level Perspective on the Role of Rents in the Rise in Inequality.” In Toward a Just Society: Joseph Stiglitz and Twenty-First Century Economics, edited by Martin Guzman. New York: Columbia University Press, 89–110.

Furman, Jason. 2025. September 26. 92% of GDP Growth. X (formerly Twitter). https://x.com/jasonfurman/status/1971995367202775284

Gaffney, Jonathan. 2024. "Judicial Review Under the Administrative Procedure Act (APA)." Congressional Research Service Legal Sidebar LSB10558. https://crsreports.congress.gov/product/pdf/LSB/LSB10558.

Galetovic, Alexander, and Stephen Haber. 2017. "The Fallacies of Patent Holdup Theory." Journal of Competition Law and Economics 13(1): 1–44.

Galetovic, Alexander, Stephen Haber, and Lew Zaretzki. 2024. “Cellular SEP Royalties and 5G: What Should Competition Policy Be?” In 5G and Beyond: Intellectual Property and Competition Policy in the Internet of Things, edited by Jonathan M. Barnett and Seán M. O’Connor. Cambridge: Cambridge University Press, 53–78.

Galetovic, Alexander, Stephen Haber, and Ross Levine. 2018. "An Empirical Examination of Patent Holdup." Journal of Competition Law and Economics 11(3): 549-578.

Galetovic, Alexander. 2021. “Patents in the History of the Semiconductor Industry: The Ricardian Rent Hypothesis.” In The Battle Over Patents, edited by Naomi Lamoreaux and Stephen Haber. Oxford: Oxford University Press: 27–68.

Gans, Joshua, David Hsu, and Scott Stern. 2002. "When Does Start-Up Innovation Spur the Gale of Creative Destruction?" RAND Journal of Economics 33(4): 571-586.

GAO. 1994. General Agreement on Tariffs and Trade: Uruguay Round Final Act Should Produce Overall U.S. Economic Gains. GAO. GGD-94-83B. https://www.govinfo.gov/content/pkg/GAOREPORTS-GGD-94-83B/html/GAOREPORTS-GGD-94-83B.htm

GAO. 1998. Technology Transfer: Administration of the Bayh-Dole Act by Research Universities. GAO/RCED-98-126. GAO. https://www.gao.gov/assets/rced-98-126.pdf

GAO. 2003. Electronic Rulemaking: Efforts to Facilitate Public Participation Can Be Improved. GAO-03-901. https://www.govinfo.gov/app/details/GAOREPORTS-GAO-03-901

Gardizy, Anissa, and Qianer Liu. 2025. “Google Convinces OpenAI to Use TPU Chips in Win Against Nvidia.” The Information. https://www.theinformation.com/articles/google-convinces-openai-use-tpu-chips-win-nvidia?rc=tuorsb

Gardner, Bruce. 2002. American Agriculture in the Twentieth Century. Cambridge: Harvard University Press.

Gates, Bill. 1976. “An Open Letter to Hobbyists.” Homebrew Computer Club Newsletter, no. 2.1 (January): 2. https://commons.wikimedia.org/wiki/File:Bill_Gates_Letter_to_Hobbyists_ocr.pdf

Gatford, Molly. 2023. “Conversational AI Market Report 2025-29.” Juniper Research. https://www.juniperresearch.com/research/telecoms-connectivity/communication-services/conversational-ai-research-report/

Gavil, Andrew, William Kovacic, and Jonathan Baker. (2014). Antitrust Law in Perspective. 3rd ed. St. Paul: West Academic Publishing.

Gawer, Annabelle, and Michael Cusumano. 2002. Platform Leadership. Boston, MA: Harvard Business School Press.

Gehan, Ann. 2025. “Amazon Locks Down Against Google’s AI Shopping Agents.” The Information. https://www.theinformation.com/articles/amazon-locks-googles-ai-shopping-agents

Gemini Team Google. 2023. Gemini. arXiv preprint arXiv:2312.11805. https://arxiv.org/abs/2312.11805

Gentry, Craig. 2009. "Fully Homomorphic Encryption Using Ideal Lattices." Proceedings of the 41st Annual ACM Symposium on Theory of Computing. New York: ACM, 169–78.

Geroski, Paul. 2000. "Models of Technology Diffusion." Research Policy 29(4-5): 603-625.

Gerstenzang, James. 1993. Senate Approves NAFTA on 61-38 Vote: Trade. https://www.latimes.com/archives/la-xpm-1993-11-21-mn-59485-story.html

Gerstle, Gary. 2022. The Rise and Fall of the Neoliberal Order. Oxford: Oxford University Press.

Gervase, Richard, Michael Renaud and Tinny Song. 2020. "Ninth Circuit Reverses FTC Win in FTC v. Qualcomm, Finding No Antitrust Violations from Qualcomm's Licensing of its Standard-

Essential Patents." Mintz Insights. https://www.mintz.com/insights-center/viewpoints/2231/2020-08-13-ninth-circuit-reverses-ftc-win-ftc-v-qualcomm-finding-no

Gestalt Diagnostics. 2025. “Gestalt Demonstrates Advanced Interoperability and Standards Compliance at 2025 DICOM WG-26 Connectathon.” Company news release. https://www.gestaltdiagnostics.com/gestalt-demonstrates-advanced-interoperability-and-standards-compliance-at-2025-dicom-wg-26-connectathon

Geva, Benjamin. 2020. Electronic Payments: Guide on Legal and Regulatory Reforms and Best Practices for Developing Countries. Osgoode Legal Studies Research Paper. https://ssrn.com/abstract=3631155

Gheller, Jonathan. 2015. “Introducing On This Day: A New Way to Look Back at Photos and Memories on Facebook.” Facebook Newsroom. https://about.fb.com/news/2015/03/introducing-on-this-day-a-new-way-to-look-back-at-photos-and-memories-on-facebook/

Gibson, Bryan. 2015. Electronic Signatures: Review and Analysis. Kentucky Transportation Center Technical Assistance Report, University of Kentucky. https://uknowledge.uky.edu/cgi/viewcontent.cgi?article=1003&context=ktc_technicalassistancereports

Gilbert, Ben and David Rosenthal. 2024. How ARM Became the World’s Default Chip Architecture (with ARM CEO Rene Haas). ACQ2 by Acquired, podcast audio. https://www.acquired.fm/episodes/how-arm-became-the-worlds-default-chip-architecture-with-arm-ceo-rene-haas

Gilbert, Richard, Carl Shapiro, Louis Kaplow, and Robert Gertner. 1997. "Antitrust issues in the licensing of intellectual property: The nine no-no's meet the nineties." Brookings Papers on Economic Activity. Microeconomics 1997: 283-349.

Gilbert, Richard, and Steven Sunshine. 1995. Incorporating Dynamic Efficiency Concerns in Merger Analysis: The Use of Innovation Markets. Antitrust Law Journal 63(2): 569-601.

Gilhousen, Klein, Irwin Jacobs, Roberto Padovani, Andrew Viterbi, Lindsay Weaver Jr., and Charles Wheatley III. 1991. "On the Capacity of a Cellular CDMA System." IEEE Transactions on Vehicular Technology 40(2): 303-312.

Gilhousen, Klein, Irwin Jacobs, Andrew Viterbi, Charles Wheatley III, Roberto Padovani, and Lindsay Weaver Jr. 1992. System and Method for Generating Signal Waveforms in a CDMA Cellular Telephone System (U.S. Patent No. 5,103,459). USPTO. https://ptacts.uspto.gov/ptacts/public-informations/petitions/1471034/download-documents?artifactId=u1dPKxpYy6ShiP9yeA_0OfKLmaBu-U6jVd22CDZ9LrOyCkVzpKhtdYw

Gilhousen, Klein, Irwin Mark Jacobs, Paul Padovani, Andrew Viterbi, Louis Weaver, and Charles Wheatley. 1987. Spread spectrum wireless PBX (U.S. Patent No. 4,672,658). USPTO. https://patents.justia.com/patent/4672658

Gillespie, Tarleton. 2018. Custodians of the Internet. New Haven, CT: Yale University Press.

Gochberg, Will and Victor Menaldo. 2022. To rent or not to rent? Mechanics, causes and consequences of Ricardian and Quasi-rents in the oil industry. Resources Policy 78(September): 102826.

Golden, Jeffrey. 2012. “The Single Agreement Concept in Derivatives Law.” Capital Markets Law Journal 7(1): 1–23.

Goldfarb, Charles. 2007. Telecommunications Act: Competition, Innovation, and Reform. CRS Report RL33034. https://www.everycrsreport.com/reports/RL33034.html

Goldin, Claudia, and Lawrence Katz. 2008. The Race between Education and Technology. Cambridge, MA: Harvard University Press.

Goldstone, Jack. 2002. "Efflorescences and Economic Growth in World History: Rethinking the 'Rise of the West' and the Industrial Revolution." Journal of World History 13(2): 323–389.

Gomes, Ben. 2017. Our latest quality improvements in Search. The Keyword (Google). https://blog.google/products/search/our-latest-quality-improvements-search/

Gonzalez. 2023. Gonzalez v. Google LLC. No. 21-1333. US Supreme Court, May 18. https://supreme.justia.com/cases/federal/us/598/21-1333/

Good, I.J. (Irving John). 1965. “Speculations Concerning the First Ultraintelligent Machine.” Advances in Computers 6: 31–88.

Goode, Bryan. 2025. From systems of record to systems of action: Dynamics 365, agentic business applications for the frontier. Microsoft Dynamics 365 Blog. https://www.microsoft.com/en-us/dynamics-365/blog/business-leader/2025/10/21/from-systems-of-record-to-systems-of-action-dynamics-365-agentic-business-applications-for-the-frontier/

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. Cambridge, MA: MIT Press.

Google Cloud. 2024. MLOps: Continuous Delivery and Automation Pipelines in Machine Learning. https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Google Cloud. 2025a. Overview of getting inferences on Vertex AI. https://docs.cloud.google.com/vertex-ai/docs/predictions/overview

Google Cloud. 2025b. AI and Machine Learning APIs Overview. Google Cloud Documentation. https://cloud.google.com/products/ai

Google. 2025b. A New Era of Intelligence with Gemini 3. The Keyword. https://blog.google/products/gemini/gemini-3/

Google. n.d. “How the Google Ads Auction Works.” Google Ads Help Center. https://support.google.com/google-ads/answer/6366577?hl=en.

Gordon, Robert. 2000. "Does the 'New Economy' Measure Up to the Great Inventions of the Past?" Journal of Economic Perspectives 14(4): 49-74.

Gordon, Robert. 2016. The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War. Princeton, NJ: Princeton University Press.

Goswami, Rohan, and Kif Leswing. 2024. Intel stock jumps on plan to turn foundry business into subsidiary and allow for outside funding. CNBC. https://www.cnbc.com/2024/09/16/intel-stock-jumps-on-plan-to-turn-foundry-business-into-subsidiary.html

Gottschalk. 1972. Gottschalk v. Benson and Tabb. No. 71-485. US Supreme Court, November 20. https://supreme.justia.com/cases/federal/us/409/63/

Gowa, Joanne. 1983. Closing the Gold Window. Ithaca: Cornell University Press.

Grace, Katja, et al. 2024. Thousands of AI Authors on the Future of AI (2023 Expert Survey on Progress in AI). AI Impacts. https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

Graetz, Georg, and Guy Michaels. 2018. "Robots at Work." The Review of Economics and Statistics 100(5): 753–768.

Graham, Stuart, and Saurabh Vishnubhakat. 2013. "Of Smart Phone Wars and Software Patents." Journal of Economic Perspectives 27(1): 67–86.

Gravelle, Jane. 2014. Bonus Depreciation: Economic and Budgetary Issues. CRS Report R43432. https://www.congress.gov/crs_external_products/R/PDF/R43432/R43432.7.pdf

Grbovic, Mihajlo, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Ricardo Baeza-Yates. 2016. “Scalable Semantic Matching of Queries to Ads in Sponsored Search Advertising.” Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’16). https://arxiv.org/pdf/1607.01869.

Green, Alastair, Andrea Del Miglio, Humayun Tai, Jesse Noffsinger, Mark Patel, Marc Sorel, and Pankaj Sachdeva. 2025. What Is a Data Center? McKinsey Explainer, McKinsey & Company. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-a-data-center#/

Greenstein, Shane. 2015. How the Internet Became Commercial. Princeton, NJ: Princeton University Press.

Griffith, Erin. 2025. "In A.I. Boom, Venture Capital Firms Are Raising Loads More Money." The New York Times. https://www.nytimes.com/2025/12/15/technology/ai-venture-capital-big-funds.html

Griliches, Zvi. 1992. “The search for R&D spillovers.” Scandinavian Journal of Economics 94 (Supplement): S29–S47.

Griliches, Zvi. 1957. “Hybrid Corn: An Exploration in the Economics of Technological Change.” Econometrica 25(4): 501–522.

Grindley, Peter, and David Teece. 1997. “Managing Intellectual Capital: Licensing and Cross-Licensing in Semiconductors and Electronics.” California Management Review 39(2): 8–41.

Grinnell. 1966. United States v. Grinnell Corp., No. 73. US Supreme Court, June 13. https://supreme.justia.com/cases/federal/us/384/563/

Guenther, Gary. 2022. Research Tax Credit. CRS Report RL31181. https://www.congress.gov/crs-product/RL31181

Gundlach, Hans, Jayson Lynch, Matthias Mertens, and Neil Thompson. 2025. The price of progress: Algorithmic efficiency and the falling cost of AI inference arXiv preprint: arXiv2511.23455. https://arxiv.org/html/2511.23455v1

Guo, Daya, et al. 2025. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Large-Scale Reinforcement Learning.” DeepSeek-AI. arXiv preprint arXiv:2501.12948. https://arxiv.org/abs/2501.12948

Gupta, Kirti. 2015. "Technology Standards and Competition in the Mobile Wireless Industry." George Mason Law Review 22: 865–893.

Gupta, Nikhil. 2025. MLOps in the Cloud-Native Era — Scaling AI/ML Workloads with Kubernetes and Serverless Architectures. Cloud Native Now. https://cloudnativenow.com/topics/cloudnativedevelopment/kubernetes/mlops-in-the-cloud-native-era-scaling-ai-ml-workloads-with-kubernetes-and-serverless-architectures/

Gurman, Mark. 2024. Gurman, Mark. 2024. “Apple to Power AI Servers With In-House Server Chips This Year.” Bloomberg. https://www.bloomberg.com/news/articles/2024-05-09/apple-to-power-ai-features-with-in-house-server-chips-this-year

Gutiérrez, Germán, and Thomas Philippon. 2018. “Ownership, Concentration, and Investment.” AEA Papers and Proceedings 108 (May): 432–37.

H2O.ai. 2025. Use AI to Assist Trading Strategies. https://h2o.ai/solutions/use-case/trading-strategies/

Haber, Stephen and Naomi Lamoreaux. 2021. “Introduction.” In The Battle Over Patents: Historical Perspectives on Current Debates, eds., Stephen Haber and Naomi Lamoreaux. Oxford: Oxford University Press: 1-26.

Haber, Stephen. 2016. "Patents and the Wealth of Nations." George Mason Law Review 23(4): 811–835.

Haber, Stephen and Naomi Lamoreaux. 2021. “Introduction.” In Patents, Inventors, and Politics: Historical Perspectives on Current Debates, eds., Stephen Haber and Naomi Lamoreaux. Oxford: Oxford University Press, 1-26.

Haggard, Stephan, and Robert Kaufman. 1992. The Politics of Economic Adjustment: International Constraints, Distributive Conflicts, and the State. Princeton, NJ: Princeton University Press.

Hall, Bronwyn H. 1996. “The private and social returns to research and development.” In Technology, R&D, and the Economy, edited by Bruce Smith and Carl Barfield. Washington, DC: Brookings Institution, 140-183.

Hall, Bronwyn, and Rosemarie Ham Ziedonis. 2001. The Patent Paradox Revisited. RAND Journal of Economics 32(1): 101-128.

Hall, Bronwyn. 2002. "The Financing of Research and Development." Oxford Review of Economic Policy 18(1): 35–51.

Hall, Robert. 1988. "The Relation between Price and Marginal Cost in U.S. Industry." Journal of Political Economy 96(5): 921-947.

Hall, Robert. 2018. “New Evidence on the Markup of Prices Over Marginal Costs and the Role of Mega-Firms in the U.S. Economy.” NBER Working Paper No. 24574. Cambridge, MA: National Bureau of Economic Research.

Hamid, Jai. 2025. Visa, Mastercard build payment rails as AI agents ready to shop and pay autonomously. Cryptopolitan. https://www.cryptopolitan.com/visa-mastercard-build-payments-ai-agents/

Hammond, George, Tabby Kinder, and Antoine Gara. 2025. “OpenAI Forges $12bn Contract with CoreWeave.” Financial Times. https://www.ft.com/content/4b52fdbb-ca8e-4208-bb99-f1e7f9313863

Handler, Milton. 1957. “Annual review of antitrust developments.” Record of the Association of the Bar of the City of New York 12: 185–202.

Handshake. 2025. State of the Graduate: Class of 2025. Handshake. https://joinhandshake.com/themes/handshake/dist/assets/downloads/network-trends/class-of-2025.pdf

Harbridge-Yong, Laurel. 2015. Is Bipartisanship Dead? Policy Agreement and Agenda-Setting in the House of Representatives. New York: Cambridge University Press

Harris, Barry, and Joseph Simons. 1989. "Focusing Market Definition." Research in Law and Economics 12: 207–26.

Harvey, David. 2005. A Brief History of Neoliberalism. Oxford: Oxford University Press.

Haskel, Jonathan, and Stian Westlake. 2017. Capitalism without Capital: The Rise of the Intangible Economy. Princeton, NJ: Princeton University Press.

Hauck, Ronny. 2015. "FTC v. Google." In Competition on the Internet, edited by Gintarė Surblytė. Berlin: Springer, 83-108.

Hausman, Jerry, and J. Gregory Sidak. 1999. "Did Mandatory Unbundling Achieve Its Purpose? Empirical Evidence from Five Countries." Federal Communications Law Journal 52(1): 3–81.

Hausman, Jerry. 1997. “Valuing the Effect of Regulation on New Services in Telecommunications.” Brookings Papers on Economic Activity: Microeconomics 1997: 1–38.

Hausmann, Ricardo, and Dani Rodrik. 2003. "Economic Development as Self-Discovery." Journal of Development Economics 72(2): 603-633.

Hawley, Ellis. 1966. The New Deal and the problem of monopoly: A study in economic ambivalence. Princeton, NJ: Princeton University Press.

Hayes, Connor. 2017. “More Ways to Share With the Facebook Camera.” Facebook Newsroom (Meta). https://about.fb.com/news/2017/03/more-ways-to-share-with-the-facebook-camera/

Hazlett, Thomas W. 1998. “Economic and political consequences of the 1996 Telecommunications Act.” Journal of Regulation and Economic Change 11 (1): 21–42.

HealthIT.gov. 2025. "Health Information Technology Standards and Certification." Office of the National Coordinator for Health Information Technology. https://www.healthit.gov/

Heaven, Will. 2025. “The great AI hype correction of 2025.” MIT Technology Review. https://www.technologyreview.com/2025/12/15/1129174/the-great-ai-hype-correction-of-2025

Helpman, Elhanan, and Manuel Trajtenberg. 1996. Diffusion of General Purpose Technologies. NBER Working Paper No. 5773. Cambridge, MA: National Bureau of Economic Research.

Henderson, Rebecca, Adam Jaffe, and Manuel Trajtenberg. 1998. “Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965–1988.” The Review of Economics and Statistics 80(1): 119–127.

Hendrycks, Dan, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2021. “The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization.” arXiv preprint arXiv:2006.16241. https://arxiv.org/abs/2006.16241

Hennessy, John and David Patterson. 2019. Computer Architecture: A Quantitative Approach. 6th ed. San Francisco: Morgan Kaufmann.

Henry, Alan. 2011. “Facebook Is Tracking Your Every Move on the Web; Here's How to Stop It.” Lifehacker. https://lifehacker.com/facebook-is-tracking-your-every-move-on-the-web-heres-5843969.

Henshall, William. 2023. “E.U.’s AI Regulation Could Be Softened After Pushback from Biggest Members.” Time. https://time.com/6338602/eu-ai-regulation-foundation-models/

Herrick. 2018. Herrick v. Grindr, LLC. No. 1:17-cv-01747. US District Court, S.D. New York, March 27. https://casetext.com/case/herrick-v-grindr-llc-1

Hestness, Joel, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, Yanqi Zhou. 2017. “Deep Learning Scaling Is Predictable, Empirically.” arXiv preprint arXiv:1712.00409. https://arxiv.org/abs/1712.00409

Hiller, Jennifer. 2025. "AI Data Centers, Desperate for Electricity, Are Building Their Own Power Plants." The Wall Street Journal. https://www.wsj.com/business/energy-oil/ai-data-centers-desperate-for-electricity-are-building-their-own-power-plants-291f5c81

hiQ Labs. 2022. hiQ Labs, Inc. v. LinkedIn Corp. No. 17-16783. US Court of Appeals, Ninth Circuit, April 18. https://cdn.ca9.uscourts.gov/datastore/opinions/2022/04/18/17-16783.pdf

HJC. 1980. Patent and Trademark Law Amendments Act of 1980. House Report No. 96-1307. 96th Congress, 2nd Session. Washington, DC: U.S. Government Printing Office.

HJC. 2020. Investigation of Competition in Digital Markets: Majority Staff Report and Recommendations. Washington, DC: U.S. Government Printing Office.

HL7 International. 2025. “FHIR Overview.” https://www.hl7.org/fhir/

Hoffman, Sarah. 2025. "Generative AI and the Future of Entry-Level Jobs." AlphaSense. https://www.alpha-sense.com/blog/trends/generative-ai-entry-level-jobs/

Hoffmann, Jordan, et al. 2022. "Training Compute-Optimal Large Language Models." arXiv preprint arXiv:2203.15556. https://arxiv.org/abs/2203.15556

Hofstadter, Richard 1955. The Age of Reform: From Bryan to F.D.R. New York: Alfred A. Knopf.

Holma, Harri, and Antti Toskala. 2012. LTE for UMTS: Evolution to LTE-Advanced. Chichester: John Wiley and Sons.

Holmes, Aaron, and Aaron Tilley. 2025. "Developers Eye Apple's Models for 'Invisible' AI Features." The Information. https://www.theinformation.com/articles/developers-eye-apples-models-invisible-ai-features

Holmes, Aaron, and Sri Muppidi. 2025. “OpenAI Takes a Page from Palantir, Doubles Down on Consulting Services.” The Information. www.theinformation.com/articles/openai-takes-page-palantir-doubles-consulting-services

Holmes, Thomas. 2001. "Bar Codes Lead to Frequent Deliveries and Superstores." The RAND Journal of Economics 32(4): 708–725.

Hounshell, David. 1984. From the American System to Mass Production, 1800-1932. Baltimore: Johns Hopkins University Press.

Hovenkamp, Herbert. 2009. The Neal Report and the Crisis in Antitrust. University of Iowa Legal Studies Research Paper No. 09-09.

Hovenkamp, Herbert. 2018. “The Rule of Reason.” Florida Law Review 70(1): 47–118.

Hovenkamp, Herbert. 2005. The Antitrust Enterprise: Principle and Execution. Cambridge, MA: Harvard University Press.

Hoyt, Bradley, Glenn Hinton, Andrew Glew, and Subramanian Natarajan. 1996. Branch target buffer for dynamically predicting branch instruction outcomes using a predicted branch history (U.S. Patent No. 5,584,001). USPTO. https://patents.google.com/patent/US5584001A

Hu, Krystal, Kenrick Cai, and Stephen Nellis. 2025. “Google works to erode Nvidia’s software advantage with Meta’s help.” Reuters. https://www.reuters.com/business/google-works-erode-nvidias-software-advantage-with-metas-help-2025-12-17/

Huang, David, Avidan Shah, Alexandre Araujo, David Wagner, and Chawin Sitawarin. 2025. "Stronger Universal and Transferable Attacks by Suppressing Refusals." NAACL 2025 (Long Papers). https://aclanthology.org/2025.naacl-long.302.pdf.

Huang, Raffaele, and Liza Lin. 2025. “Chinese AI Companies Dodge U.S. Chip Curbs by Flying Suitcases of Hard Drives Abroad.” Wall Street Journal. https://www.wsj.com/tech/china-ai-chip-curb-suitcases-7c47dab1

Hubbard, Patrick. 2018. "The Artificial Intelligence 'Black Box' and the Failure of Intent and Causation." Harvard Journal of Law and Technology 31(2): 890-938.

Huescas, Jose Carlos. 2024. Unleashing the Power of AI: MLPerf Benchmarking Outcomes Show. Lenovo Press. https://lenovopress.lenovo.com/lp1969-unleashing-the-power-of-ai-at-the-edge-mlperf-benchmarking-outcomes

Hugging Face. 2025a. Inference Endpoints. Hugging Face Documentation. https://huggingface.co/docs/inference-endpoints

Hugging Face. 2025b. deepseek-ai/DeepSeek-R1 (model card; accessed Nov. 2025). https://huggingface.co/deepseek-ai/DeepSeek-R1

Hugging Face. 2025c. meta-llama/Llama-3.1-8B-Instruct (model card; accessed Nov. 2025). https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Hughes, Thomas. 1983. Networks of power: Electrification in Western society, 1880–1930. Baltimore: Johns Hopkins University Press.

Humphrey's Executor 1935. Humphrey's Executor v. United States, No. 667, U.S. Supreme Court, May 27. https://supreme.justia.com/cases/federal/us/295/602/

Hunter, Matt. 2016. "Instagram launches 'Stories,' a product to take on Snapchat." CNBC. https://www.cnbc.com/2016/08/02/instagram-launches-stories-a-product-to-take-on-snapchat.html

Hussein, Ahmed, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. "Imitation learning: A survey of learning methods." ACM Computing Surveys 50 (2): 21.

Hwang, Tim. 2020. Subprime Attention Crisis. New York: FSG Originals.

IBM. 1969. United States v. International Business Machines Corporation. 1969. Civil Action No. 69 Civ. 200 (S.D.N.Y.). https://www.justice.gov/atr/case-document/united-states-memorandum-1969-case

IBM. 1982. United States v. International Business Machines Corp., No. 69 Civ. 200 (S.D.N.Y.), Stipulation of Dismissal, Jan. 8. https://law.justia.com/cases/federal/district-courts/FSupp/539/473/2151775

IBM Research. 2025. Agent Communication Protocol (ACP). https://research.ibm.com/projects/agent-communication-protocol

IDC. 2023. Worldwide Global DataSphere Forecast, 2023–2027: It's an AI Data World. IDC. https://www.marketresearch.com/IDC-v2477/Worldwide-IDC-Global-DataSphere-Forecast-33986214/

Iden, Virginius Gilmore. 1913. The Federal Reserve Act of 1913: History and Digest. FRASER. https://fraser.stlouisfed.org/title/federal-reserve-act-1913-962/fulltext

IEA. 2025. Energy and AI. Paris: IEA. https://www.iea.org/reports/energy-and-ai

IETF. 2022. "Robots Exclusion Protocol." RFC 9309. https://datatracker.ietf.org/doc/html/rfc9309

IFR. 2020. “Record 2.7 Million Robots Work in Factories Around the Globe.” IFR Press Release. https://ifr.org/ifr-press-releases/news/record-2.7-million-robots-work-in-factories-around-the-globe

Illinois Tool Works Inc. 2006. Illinois Tool Works Inc. et al. v. Independent Ink, Inc., No. 04-1329, U.S. Supreme Court, March 1. https://scholar.google.com/scholar_case?case=12062423076831474839

InfoQ. 2025. "OpenAI and Anthropic Donate AGENTS.md and Model Context Protocol to New Agentic AI Foundation." InfoQ. https://www.infoq.com/news/2025/12/agentic-ai-foundation/

Innovation Alliance. 2022. Public Submission on SEPs. https://innovationalliance.net/wp-content/uploads/2022/02/2022-0204-Innovation-Alliance-SEP-Public-Submission.pdf.

InsideHPC. 2024. Nvidia Announces Installations of 9 Grace Hopper-Powered Supercomputers. InsideHPC. https://insidehpc.com/2024/05/nvidia-announces-installations-of-9-grace-hopper-powered-supercomputers/

Insilico Medicine. 2023. “ReLEHF: Reinforcement Learning with Expert Human Feedback to Advance Pharmaceutical Generative AI.” Insilico Medicine Blog. https://insilico.com/blog/relehf

Intel Corp. 2024. Form 10-K: Annual report for the fiscal year ended December 30, 2023. U.S. SEC. https://www.intc.com/filings-reports/annual-reports

Ip, Greg. 2025. “How Global Trade Could Survive Trump’s Tariffs.” The Wall Street Journal. https://www.wsj.com/economy/trade/how-global-trade-could-survive-trumps-tariffs-24d74d08?mod=Searchresults_pos1&page=1.

Irwin, Douglas. 1996. "The Smoot-Hawley Tariff: A Quantitative Assessment." NBER Working Paper No. 5509. Cambridge, MA: National Bureau of Economic Research.

Irwin, Douglas. 2013. "The Nixon Shock after Forty Years." World Trade Review 12(1): 29–56.

Irwin, Douglas, and Peter Klenow. 1996. "High-Tech R&D Subsidies: Estimating the Effects of Sematech." Journal of International Economics 40(3–4): 323–344.

ISA. 2025. “ISA95, Enterprise-Control System Integration.” https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa95

Isaac. 2025. How much does it cost to manufacture a mobile phone worth more than 1000 euros? Android Guias. https://en.androidguias.com/How-much-does-it-cost-to-manufacture-a-mobile-phone-worth-more-than-1000-euros%3F/

Isaacson, Walter 2014. The Innovators: How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution. New York: Simon and Schuster.

ISDA. 2002. ISDA Master Agreement. New York: ISDA.

ISO/IEC JTC 1/SC 42. 2024. Artificial Intelligence. ISO/IEC JTC 1/SC 42. https://jtc1info.org/sd-2-history/jtc1-subcommittees/sc-42

ITC. 1993 Potential Impact on the U.S. Economy and Selected Industries of the North American Free-Trade Agreement. ITC Publication 2596. https://www.usitc.gov/publications/332/pub2596.pdf

ITC. 1978. Annual Report 1978. ITC. https://www.usitc.gov/publications/annualreport/pub982.pdf

ITC. 1994. The Uruguay Round of Multilateral Trade Negotiations: Summary of the Final Act. ITC Publication 2790. https://www.usitc.gov/publications/332/pub2790.pdf

Ivanov, Dmitry. 2021. "Supply Chain Resilience: Modelling, Management, and Control." International Journal of Production Research 59(8): 2535–2551.

Jaffe, Adam, Manuel Trajtenberg, and Rebecca Henderson. 1993. “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations.” Quarterly Journal of Economics 108(3): 577–598.

Janakiram, MSV. 2025. "Nvidia Wants To Build A Planet-Scale AI Factory With DGX Cloud Lepton." Forbes. https://www.forbes.com/sites/janakirammsv/2025/06/13/nvidia-wants-to-build-a-planet-scale-ai-factory-with-dgx-cloud-lepton/

Janc, Victor, and Steven Jasper. 1988. Digital Global Positioning System Receiver (US Patent No. 4,785,463). USPTO. https://patents.google.com/patent/US4785463A/en

Jefferson Parish Hospital District No. 2. 1984. Jefferson Parish Hospital District No. 2 et al. v. Hyde. No. 82-1031. US Supreme Court, March 27. https://supreme.justia.com/cases/federal/us/466/2

Jiang, Albert, et al. 2023. “Mistral 7B.” arXiv preprint arXiv:2310.06825. https://arxiv.org/abs/2310.06825

Jie, Yang. 2025. Intel, SoftBank Discussed a Sale. Now, a Cash Infusion Will Accelerate American AI. The Wall Street Journal. https://www.wsj.com/tech/intel-gets-2-billion-infusion-from-softbank-accelerating-ai-push-in-u-s-2129b3c4

Jin, Berber, Peter Rudegeair, and Matt Wirz. 2025. “Musk Allies to Raise Up to $12 Billion for xAI Chips as Startup Burns Through Cash.” Wall Street Journal. https://www.wsj.com/tech/ai/elon-musk-x-ai-funding-feecede1

John Mac Ghlionn. 2025. “A blue-collar bloodbath is inevitable.” The Hill. https://thehill.com/opinion/technology/5445562-ai-threat-to-skilled-trades/

Johnson, Jennifer and George Hay. 2025. “Europe will struggle to get Big Tech off its cloud.” Reuters Breakingviews. https://www.reuters.com/commentary/breakingviews/europe-will-struggle-get-big-tech-off-its-cloud-2025-06-26/

Jon Peddie Research. 2024. Summary report on the worldwide total GPU market. JPR. https://www.jonpeddie.com/store/summary-report-on-the-worldwide-total-gpu-market/

Jonas, Adam, et al. 2024. Humanoids: Investment Implications of Embodied AI. Morgan Stanley Blue Paper. https://www.futuremanagementgroup.com/wp-content/uploads/240626-Humanoid-Robots-Morgan-Stanley.pdf

Jonas, Adam, William Tackett, Sheng Zhong, Daniela Haigian, and Elizabeth Tso. 2025. The Humanoid 100: Mapping the Humanoid Robot Value Chain. Morgan Stanley.

Joosse, Tess. 2022. “December 1945: The ENIAC Computer Runs Its First, Top-Secret Program.” American Physical Society: APS News/This Month in Physics History. https://www.aps.org/apsnews/2022/11/eniac-first-top-secret-program

Jorgenson, Dale. 2001. "Information Technology and the U.S. Economy." American Economic Review 91(1): 1-32.

Jorgenson, Dale, and Kevin Stiroh. 2000. "Raising the Speed Limit: U.S. Economic Growth in the Information Age." Brookings Papers on Economic Activity 2000(1): 125-211.

Jouppi, N. P., et al. 2017. In-datacenter performance analysis of a tensor processing unit. In Proceedings of the 44th annual international symposium on computer architecture: 1-12.

Jovanovic, Boyan, and Peter Rousseau. 2005. "General Purpose Technologies." In Handbook of Economic Growth, Volume 1, Part B. Edited by Philippe Aghion and Steven Durlauf. Amsterdam: Elsevier, 1181–1224.

Jumper, John, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature (5967873): 583–589.

Junarkar, Sandeep. 2003. "HP seals P&G outsourcing deal."

ZDNet. https://www.zdnet.com/article/hp-seals-p-g-outsourcing-deal/

Kadrey et al. 2023. Kadrey et al. v. Meta Platforms Inc., No. 3:23-cv-03417. US District Court for the Northern District of California, July 7.

https://www.courtlistener.com/docket/67569326/kadrey-v-meta-platforms-inc

Kagan, Julia. 2024. Lump of Labor Fallacy: Definition and How It Works. Investopedia.

https://www.investopedia.com/terms/l/lump-of-labour-fallacy.asp

Kahn, Jeremy. 2025. “OpenAI Builds Apps into ChatGPT, in a Bold Bid to Make AI the Universal Interface.” Fortune. https://fortune.com/2025/10/07/openais-chatgpt-apps-universal-interface-platform-shift-amd-deal-eye-on-ai

Kang, Cecilia, Cade Metz, and Stuart Thompson. 2024. “Four Takeaways on the Race to Amass Data for A.I.” New York Times. https://www.nytimes.com/2024/04/06/technology/ai-data-tech-takeaways.html.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. arXiv preprint: arXiv:2001.08361 https://doi.org/10.48550/arXiv.2001.08361

Kaplow, Louis. 1987. "Antitrust, Law and Economics, and the Courts." *Law and Contemporary Problems* 50(4): 181–216.

Kaplow, Louis. 2010. "Why (Ever) Define Markets?" Harvard Law Review 124(2): 437–517.

Karlsson, Kasper. 2023. "Arm's Smartphone Monopoly: IPO and Beyond." Quartr Insights. https://quartr.com/insights/company-research/arm-s-smartphone-monopoly-setting-the-stage-for-its-ipo

Katsurashima, Wataru, Sid Nag, and Evan Zeng. 2024. Competitive Behaviors of Hyperscale Providers in the AI Infrastructure Market. Gartner Research. https://www.gartner.com/document/competitive-behaviors-hyperscale-providers-ai-infrastructure-market

Katz, Michael, and Jonathan Sallet. 2018. Multisided Platforms and Antitrust Enforcement. Yale Law Journal 127(7): 2142-2175.

Katz, Michael, and Carl Shapiro. 1994. "Systems Competition and Network Effects." Journal of Economic Perspectives 8(2): 93–115.

Katz, Michael, and Carl Shapiro. 2003. "Critical Loss: Let's Tell the Whole Story." *Antitrust* (Spring): 49–56.

Katz, Michael, and Howard Shelanski. 2005. "'Schumpeterian' Competition and Antitrust Policy in High-Tech Markets." Competition 14(2): 47-91.

Katz, Michael, and Howard Shelanski. 2007. "Schumpeterianism and Antitrust." Villanova Law Review 52(3): 513–562.

Katz, Richard. 1998. Japan: The System That Soured: The Rise and Fall of the Japanese Economic Miracle. Armonk, NY: M.E. Sharpe.

Katz, Richard. 2025. "A 21st Century Industrial Policy: Diffusing Technology from Leading Companies to Laggards." Japan Economy Watch, Dec. 4th 2025 Entry. https://richardkatz.substack.com/p/a-21st-century-industrial-policy

Katzen, Sally. 2011. "OIRA at Thirty." Administrative Law Review 63 (Special Edition): 103–114.

Kauper, Thomas. 1981. "The Burger Court and Antitrust Law: An Unhappy Marriage." Antitrust Law Journal 50(4): 783–801

Kay, Grace. 2025. Internal xAI org chart: Names, positions, team size. Business Insider. https://www.businessinsider.com/xai-org-chart-employees-elon-musk-direct-reports-2025-3

Kaysen, Carl. 1956. United States v. United Shoe Machinery Corporation. Cambridge, MA: Harvard University Press.

Kearns, Michael, and Yuriy Nevmyvaka. 2013. "Machine Learning for Market Microstructure and High Frequency Trading." In *High Frequency Trading: New Realities for Traders, Markets and Regulators*, edited by Irene Aldridge. London: Risk Books, 67–93.

Kehayias, Alex. 2023. SaaS Revenue per Employee Benchmarks for B2B Companies. SaaS Capital. https://notes.alexkehayias.com/revenue-per-employee-benchmarks/

Kennedy, Paul. 2017. Vampire Capitalism London: Palgrave Macmillan.

Kenney, Martin, and John Zysman. 2016. “The Rise of the Platform Economy.” Issues in Science and Technology 32(3).

Kessler, Andy. 2022. Putin’s vertical empire: Power and money in Russia. The Wall Street Journal. https://www.wsj.com/articles/putins-vertical-empire-fall-ukraine-russia-business-kleptocrat-oligarchs-value-companies-11646581296

Kessler, Andy. 2025. “Netscape’s Lessons for AI Mania.” The Wall Street Journal. https://www.wsj.com/opinion/netscapes-lessons-for-ai-mania-innovation-technology-a132d8ae

Khan, Lina M. 2019. "The Separation of Platforms and Commerce." *Columbia Law Review* 119(7): 973–1095. Columbia Law Review.

Khan, Lina. 2017. "Amazon's Antitrust Paradox." Yale Law Journal 126(3): 710–805.

Khronos Group. 2025. What is SYCL? C++ programming for heterogeneous parallel computing. Khronos Group. https://www.khronos.org/sycl/

Kiela, Douwe, et al. 2021. "Dynabench: Rethinking Benchmarking in NLP." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4110–4124.

Kilpatrick, Jim, Berckman, Lindsey, Faver, Alan D., Hardin, Kate, and Sloane, Matt. 2025. Supply chain resilience. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/manufacturing/global-supply-chain-resilience-amid-disruptions.html

Kim, Sang Hui Michael. 1995. "In re Alappat: A Strict Statutory Interpretation Determining Patentable Subject Matter Relating to Computer Software?" John Marshall Journal of Information Technology and Privacy Law 13(4): 635–665.

Kincaid, Jason. 2009. "Facebook Activates 'Like' Button; FriendFeed Tires Of Sincere Flattery." TechCrunch. https://techcrunch.com/2009/02/09/facebook-activates-like-button-friendfeed-tires-of-sincere-flattery/

Kintner, Earl. 1978. The Legislative History of the Federal Antitrust Laws and Related Statutes. Chelsea House Publishers.

Kirchenbauer, Jonas, John Thickstun, Ananya Kumar, Dan Jurafsky, Percy Liang, and Tatsunori Hashimoto. 2023. A Watermark for Large Language Models. arXiv preprint: arXiv.2301.10226. https://doi.org/10.48550/arXiv.2301.10226

Kirkpatrick, David. 2010. The Facebook Effect. New York: Simon and Schuster.

Klarna. 2024. “Klarna AI Assistant Handles Two-Thirds of Customer Service Chats in Its First Month.” Press Release. https://www.prnewswire.com/news-releases/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month-302072740.html

Klein, Aaron. 2019. “Not all robots take your job, some become your co-worker.” Brookings. https://www.brookings.edu/articles/not-all-robots-take-your-job-some-become-your-co-worker/

Klein, Benjamin, Robert Crawford, and Armen Alchian. 1978. "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process." Journal of Law and Economics 21(2): 297–326.

Klein, Joel. 1998. "The Importance of Antitrust Enforcement in the New Economy." Speech delivered at the New York Law Journal Invitational Program on Antitrust, New York, NY, January 29. DoJ. https://www.justice.gov/atr/speech/importance-antitrust-enforcement-new-economy

Kline, Patrick, and Enrico Moretti. 2014. "Local Economic Development, Agglomeration Economies, and the Big Push." The Quarterly Journal of Economics 129(1): 275–331.

Klobuchar, Amy. 2024. "Klobuchar Statement Following the Supreme Court Decision to Overturn Chevron Deference." U.S. Senator Amy Klobuchar (press release). https://www.klobuchar.senate.gov/public/index.cfm/2024/6/klobuchar-statement-following-the-supreme-court-decision-to-overturn-chevron-deference.

Klonick, Kate. 2018. "The New Governors." Harvard Law Review 131(6): 1598–1670.

Klor's. 1959. Klor's Inc. v. Broadway-Hale Stores, Inc., No. 76. US Supreme Court, April 6. https://supreme.justia.com/cases/federal/us/359/207

Kober, Jens, J. Andrew Bagnell, and Jan Peters. 2013. "Reinforcement Learning in Robotics: A Survey." International Journal of Robotics Research 32(11): 1238-1274.

Kolodny, Lora and Jennifer Elias. 2024. Waymo dominated U.S. robotaxi market in 2024, but Tesla, Zoox loom. CNBC. https://www.cnbc.com/2024/12/26/waymo-dominated-us-robotaxi-market-in-2024-but-tesla-zoox-loom.html

Kosseff, Jeff. 2019. The Twenty-Six Words That Created the Internet. Ithaca, NY: Cornell University Press.

Koster, Martijn. 1994. "A Standard for Robot Exclusion." Robots Pages. http://www.robotstxt.org/orig.html

Kovach, Steve. 2023. "Twitter Users Get Error Messages as Elon Musk Outlines New Limits, Blames A.I. Data Scraping." Fortune. https://fortune.com/2023/07/01/twitter-error-messages-technical-problems-elon-musk-rate-limit-exceeded/

Kovacic, William. 2007. The Intellectual DNA of Modern U.S. Competition Law for Dominant Firm Conduct: The Chicago/Harvard Double Helix. Columbia Business Law Review 2007(1): 1-80.

Kovacic, William E., and Carl Shapiro. 2000. "Antitrust Policy: A Century of Economic and Legal Thinking." Journal of Economic Perspectives 14(1): 43–60.

Kovacic, William. 1990. "The Antitrust Paradox Revisited." Wayne Law Review 36(3): 1413–1471.

Kovacic, William. 2003. "The Modern Evolution of US Competition Policy Enforcement Norms." Antitrust Law Journal 71(2): 377–478.

Kratz, Agatha, Piper, Lauren, and Bouchaud, Juliana. 2024. China and the future of global supply chains. Rhodium Group. https://rhg.com/research/china-and-the-future-of-global-supply-chains/

Kremer, Michael. 1993. "The O-Ring Theory of Economic Development." The Quarterly Journal of Economics 108(3): 551–575.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25* (NeurIPS 2012), 1097–1105.

Krug, Sammi. 2016. “What the Reactions Launch Means for News Feed.” Facebook Newsroom. https://about.fb.com/news/2016/02/news-feed-fyi-what-the-reactions-launch-means-for-news-feed/

KSR 2007. KSR International Co. v. Teleflex, Inc. No. 04-1350. US Supreme Court, April 30. https://supreme.justia.com/cases/federal/us/550/398/

Kubeflow. 2024. Kubeflow: Machine Learning Toolkit for Kubernetes. Kubeflow Documentation. https://www.kubeflow.org/docs/

Kubernetes. 2024. Kubernetes documentation. https://kubernetes.io/docs/

Kwa, Amanda, Elizabeth Barnes, Lawrence Chan, and Ben West. 2025. Measuring AI Ability to Complete Long Tasks. arXiv preprint arXiv:2503.14499. https://arxiv.org/abs/2503.14499

Kwoka, John, and Lawrence White, eds. 2014. The Antitrust Revolution: Economics, Competition, and Policy. 6th ed. Oxford: Oxford University Press.

Kyosovska, Nicoleta and Andrea Renda. 2025. “EU Plans for AI (Giga)Factories: Sanctuaries of Innovation, or Cathedrals in the Desert?” Centre for European Policy Studies (CEPS) Working Paper. https://cdn.ceps.eu/2025/11/251027-Sanctuaries-or-Cathedrals.pdf

Laffont, Jean-Jacques, and Jean Tirole. 2000. Competition in Telecommunications. Cambridge, MA: MIT Press.

Lam Research. 2024. Form 10-K and Annual Report filings (gross margin; business description). Lam Research Corporation, Investor Relations/SEC filings portal.

https://filecache.investorroom.com/mr5ir_lamresearch2/1411/LRCX%20%28Lam%20Research%20Corporation%29%20%20%2810-K%29%202024-08-29.pdf_.pdf

Landes, David. 2008. The unbound Prometheus. Cambridge: Cambridge University Press.

Landes, William M., and Richard Posner. 1981. "Market Power in Antitrust Cases." Harvard Law Review 94(5): 937–996.

Langlois, Richard and Paul Robertson. 1995. Firms, Markets, and Economic Change: A Dynamic Theory of Business Institutions. London: Routledge.

Laricchia, Federica. 2024. "Apple: Expenditure on Research and Development 2007–2024." Statista. https://www.statista.com/statistics/273006/apple-expenses-for-research-and-development/.

Lascelles, Eric, and Josh Nye. 2025. MacroMemo – October 14 – November 3, 2025. RBC Global Asset Management. https://institutional.rbcgam.com/en/us/research-insights/article/macromemo-october-14-november-3-2025-us/detail

Lau, Johann. 2020. "Google Maps 101." The Keyword (Google blog). https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/

Lawson, Emma. 2025. AI in drug discovery: AlphaFold to clinical breakthroughs. https://www.elbuenonews.com/ai-steps-up-in-drug-discovery-real-world-breakthroughs-from-lab-to-clinic/

Layne-Farrar, Anne, Jorge Padilla, and Richard Schmalensee. 2007. "Pricing Patents for Licensing in Standard-Setting Organizations: Making Sense of FRAND Commitments." Antitrust Law Journal 74(3): 671-706.

Lazonick, William. 2009. Sustainable Prosperity in the New Economy? Business Organization and High-Tech Employment in the United States. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep learning.” Nature 521: 436–444.

Lee, Jinwoo, Fangming Wu, Wei Zhao, Mohsen Ghaffari, Linxia Liao, and David Siegel. 2014. "Prognostics and Health Management Design for Rotary Machinery Systems." *Mechanical Systems and Signal Processing* 42(1-2): 314–334.

Lee, Kevin and Shubho Sengupta. 2022. "Introducing the AI Research SuperCluster." Meta AI Engineering. https://ai.meta.com/blog/ai-rsc/

Lemley, Mark, and A. Douglas Melamed. 2013. "Missing the Forest for the Trolls." Columbia Law Review 113(8): 2117-2189.

Lemley, Mark, and Carl Shapiro. 2007. Patent Holdup and Royalty Stacking. Texas Law Review, 85(7): 1991–2049.

Lemmon. 2021. Lemmon v. Snap, Inc. No. 19-55362. US Court of Appeals, Ninth Circuit, May 4. https://law.justia.com/cases/federal/appellate-courts/ca9/19-55362/19-55362-2021-05-04.html

Leopold, Till. 2025. "Is AI Closing the Door on Entry-Level Job Opportunities?" World Economic Forum. https://www.weforum.org/stories/2025/04/ai-jobs-international-workers-day/

Lerner, Abba. 1934. "The Concept of Monopoly and the Measurement of Monopoly Power." Review of Economic Studies 1(3): 157-175.

Lerner, Josh. 1999. "The Government as Venture Capitalist." The Journal of Business 72(3): 285–318.

Lessard, Kate. 2025. The Dreamforce 2025 product announcements you can't miss. Salesforce Admins Blog. https://admin.salesforce.com/blog/2025/dreamforce-product-announcements-for-admins

Lessig, Lawrence. 2001. The Future of Ideas. New York: Vintage Books.

Leswing, Kif. 2021. "Apple's ad privacy change impact shows the power it wields over other industries." CNBC. https://www.cnbc.com/2021/11/13/apples-privacy-changes-show-the-power-it-holds-over-other-industries.html

Leswing, Kif. 2024. Nvidia's latest AI chip will cost more than $30,000, CEO says. CNBC. https://www.cnbc.com/2024/03/19/nvidias-blackwell-ai-chip-will-cost-more-than-30000-ceo-says.html

Leswing, Kif. 2025. "Apple cuts App Store fee in half for 'mini apps' that integrate more of its software." CNBC. https://www.cnbc.com/2025/11/13/apple-announces-new-program-that-cut-mini-app-fees-in-half.html

Lev, Baruch. 2001. Intangibles: Management, Measurement, and Reporting. Washington, D.C.: Brookings Institution Press.

Levi Stuart, William Ridgway David Simon, Meredith Slawe and Anita Oh. 2024. "Utah Becomes First State To Enact AI-Centric Consumer Protection Law." Skadden, Arps, Slate, Meagher and Flom LLP. https://www.skadden.com/insights/publications/2024/04/utah-becomes-first-state

Levin, Cole and Ruiz-Reyes, Federico. 2025. From seats to success: The new economics of AI pricing in 2025. Pilot Blog. https://pilot.com/blog/ai-pricing-economics-2025

Levin, Jonathan. 2011. "The Economics of Internet Markets." NBER Working Paper No. 16852. Cambridge, MA: National Bureau of Economic Research.

Levy, Ari. 2025. Elon Musk says 80% of Tesla's value will eventually come from Optimus. CNBC. https://www.cnbc.com/2025/09/02/musk-tesla-value-optimus-robot.html

Levy, David and Steve Weizer. 1985. "System Error." Regulation 9(5–6): 27-30.

Lin, Andy. 2023. "The TSMC Cost, Sell Price, and R&D Cost of Chip Foundry." Andy Lin's Long-term Stock Investment Blog. https://www.granitefirm.com/blog/us/2023/04/29/cost-of-chip-foundry/

Lin, Hesheng, Van der Plas, Geert, Sun, Xiao, Velenis, Dimitrios, Catthoor, Francky, Lauwereins, Rudy, and Beyne, Eric. 2022. Efficient Backside Power Delivery for High-Performance Computing Systems. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 30(11): 1748–1756.

Lin, Liza, Josh Chin, and Raffaele Huang. 2025. "China Is Quickly Eroding America's Lead in the Global AI Race." The Wall Street Journal https://www.wsj.com/tech/ai/artificial-intelligence-us-vs-china-03372176

Linux Foundation. 2023. *2023 State of OSPOs and OSS Initiatives*. Linux Foundation Research. https://www.linuxfoundation.org/research/sponsorship/state-of-ospos-2023

Lipscy, Phillip 2020. COVID-19 and the politics of crisis. International Organization 74(S1): E98–E127.

Lipsey, Richard, Kenneth Carlaw, and Clifford Bekar. 2005. Economic Transformations: General Purpose Technologies and Long Term Economic Growth. Oxford University Press.

Lipsky, Abbott 1981. "Current Antitrust Division Views on Patent Licensing Practices." Remarks before the American Bar Association Antitrust Section, Washington, D.C., November 5. https://www.justice.gov/atr/speech/current-antitrust-division-views-patent-licensing-practices

Litjens, Geert, et al. 2017. "A Survey on Deep Learning in Medical Image Analysis." Medical Image Analysis 42: 60–88.

Liu, Glory. 2022. Adam Smith's America. Princeton: Princeton University Press.

Liu, Qianer. 2025. “Huawei Unveils AI Chip Roadmap to Challenge Nvidia Dominance.” The Information. https://www.theinformation.com/briefings/huawei-unveils-ai-chip-roadmap-challenge-nvidia-dominance

Liu, Xing, and Price Fishback. 2019. Effects of New Deal Spending and the downturns of the 1930s on private labor markets in 1939/1940. Explorations in Economic History 71(1): 25–54.

LiveRamp. 2024. “Walled Gardens in Advertising Explained.” LiveRamp Blog. https://liveramp.com/blog/walled-gardens-advertising-explained

Lomas, Natasha. 2024. “German LLM Maker Aleph Alpha Pivots to AI Support.” TechCrunch. https://techcrunch.com/2024/09/05/german-llm-maker-aleph-alpha-pivots-to-ai-support/

Long, Clarisa. 1997. TRIPs and Intellectual Property Protection in Emerging Markets. Intellectual Property Practice Group Newsletter, The Federalist Society 3(1). https://fedsoc.org/commentary/publications/trips-and-intellectual-property-protection-in-emerging-markets

Lopatka, John. 2000. United States v. IBM: A Monument to Arrogance. Antitrust Law Journal 68(1): 145-186.

Loper Bright 2024. Loper Bright Enterprises et al. v. Raimondo, Secretary of Commerce, et al., No. 22-451, U.S. Supreme Court, June 28. https://www.supremecourt.gov/opinions/23pdf/22-451_7m58.pdf

Louviere, Jordan, David Hensher, and Joffre Swait. 2000. Stated Choice Methods: Analysis and Applications. Cambridge: Cambridge University Press.

Luhnow, David, and Tom Fairless. 2025. “Europe Is Losing.” Wall Street Journal. https://www.wsj.com/world/europe/europe-is-losing-fe179376

Luminance. 2025. Luminance: Legal-Grade AI Platform. https://www.luminance.com/

Lévêque, François, and Yann Ménière. 2008. “Licensing 3G Mobile Technology: How Did We Arrive at a Royalty Stack?” CERNA, Mines ParisTech. https://ideas.repec.org/p/cmf/wpaper/wp2007_0701.html

Ma, Wayne, and Qianer Liu. 2024. "Apple Is Working on AI Server Chips With Broadcom." The Information. https://www.theinformation.com/articles/apple-is-working-on-ai-server-chips-with-broadcom

Machado, Adrian. 2025. How tiered pricing enhances your API monetization strategy. DEV Community. https://dev.to/zuplo/how-tiered-pricing-enhances-your-api-monetization-strategy-12io

Macher, Jeffrey, and David Mowery. 2004. "Vertical Specialization and Industry Structure in High Technology Industries." In Business Strategy over the Industry Lifecycle (Advances in Strategic Management Vol. 21), edited by Joseph Baum and Anita McGahan. Bingley: Emerald Group Publishing Limited, 317–355.

MacMillan, Douglas. 2018. “Tech’s ‘Dirty Secret’: The App Developers Sifting Through Your Gmail.” The Wall Street Journal. https://www.wsj.com/articles/techs-dirty-secret-the-app-developers-sifting-through-your-gmail-1530544442

Macrotrends. n.d.a. Apple Market Cap 2020. Macrotrends. https://www.macrotrends.net/stocks/charts/AAPL/apple/market-cap#AAPL-2020

Macrotrends (n.d.b). Microsoft Market Cap 2020. Macrotrends. https://www.macrotrends.net/stocks/charts/MSFT/microsoft/market-cap#MSFT-2020

Macrotrends (n.d.c). Amazon Market Cap 2020. Macrotrends. https://www.macrotrends.net/stocks/charts/AMZN/amazon/market-cap#AMZN-2020

Macrotrends (n.d.d). Alphabet Market Cap 2020. Macrotrends. https://www.macrotrends.net/stocks/charts/GOOGL/alphabet/market-cap#GOOGL-2020

Macrotrends (n.d.e). Nvidia Market Cap 2025. Macrotrends. https://www.macrotrends.net/stocks/charts/NVDA/nvidia/market-cap#NVDA-2025-07-08

Macrotrends (n.d.f). Microsoft Market Cap 2025. Macrotrends. https://www.macrotrends.net/stocks/charts/MSFT/microsoft/market-cap#MSFT-2025-07-08

Macrotrends (n.d.g). Apple Market Cap 2025. Macrotrends. https://www.macrotrends.net/stocks/charts/AAPL/apple/market-cap#AAPL-2025-07-08

Macrotrends (n.d.h). Alphabet Market Cap 2025. Macrotrends. https://www.macrotrends.net/stocks/charts/GOOGL/alphabet/market-cap#GOOGL-2025-07-08

Macrotrends. n.d. "Historical Stock Price Data." Macrotrends. https://www.macrotrends.net/

Madden, Keith. 2025. File Types and Artificial Intelligence: Data Formats for AI Training. Hex Browser. https://www.hexbrowser.com/file-types-artificial-intelligence-data/.

Magistro, Beatrice and Victor Menaldo. 2025. How Populism Harms Prosperity: Unified Populist Rule Reduces Investment, Innovation, and Productivity. Journal of Evolutionary Economics 35(3): 553–586.

Mahaseth, Raushan Kumar. 2025. Reinforcement Learning with Human Feedback (RLHF): Shaping the Future of AI Alignment Roadmap 2025–2035. TechRxiv. https://www.techrxiv.org/users/963437/articles/1335644-reinforcement-learning-with-human-feedback-rlhf-shaping-the-future-of-ai-alignment-roadmap-2025-2035

Malik, Aisha. 2022. “Meta says Reels now makes up over 20% of the time users spend on Instagram.” TechCrunch. https://techcrunch.com/2022/04/27/meta-says-reels-now-makes-up-over-20-of-the-time-users-spend-on-instagram/

Mallinson, Keith. 2021. Royalty pricing dichotomy in 5G SEP patent pool for Open RAN Radio Units. IP Finance. http://www.ip.finance/2021/12/royalty-pricing-dichotomy-in-5g-sep.htm

Malwarebytes. 2020. Malwarebytes, Inc. v. Enigma Software Group USA, LLC. No. 19-1284. US Supreme Court, October 13. https://www.law.cornell.edu/supremecourt/text/19-1284

Mandler, Clare 2024. Three Mile Island nuclear power plant will reopen for Microsoft. NPR. https://www.npr.org/2024/09/20/nx-s1-5120581/three-mile-island-nuclear-power-plant-microsoft-ai

Manenti, Pierfrancesco. 2021. Supply Chain Strategic Planning Must Account for 5 Key Capabilities. Gartner Research. https://www.scmr.com/article/strategic_planning_must_factor_in_5_key_capabilities

Mann, Ronald, and Thomas Sager. 2007. "Patents, Venture Capital, and Software Startups." Research Policy 36(2): 193–208.

Mansfield, Edwin. 1961. "Technical Change and the Rate of Imitation." Econometrica 29(4): 741–766.

Manyika, James, Sree Ramaswamy, Jacques Bughin, Jonathan Woetzel, Michael Birshan, and Zubin Nagpal. 2017. *The Age of Analytics: Competing in a Data-Driven World*. McKinsey Global Institute. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

MAPI Foundation. 2017. "U.S. Industrial Outlook: Glimmers of Light." MAPI Foundation. https://mapifoundation.org/economic/2017/2/26/us-industrial-outlook-glimmers-of-light/

Marcus, Gary. 2025a. AI Agents have, so far, mostly been a dud. Marcus on AI (Substack). https://garymarcus.substack.com/p/ai-agents-have-so-far-mostly-been

Marcus, Gary. 2025b. “A Knockout Blow for LLMs? LLM ‘Reasoning’ Is So Cooked They Turned My Name into a Verb.” Marcus on AI (Substack). https://garymarcus.substack.com/p/a-knockout-blow-for-llms

Marcus, Gary. 2025c. “AI, Layoffs, Productivity and The Klarna Effect.” Marcus on AI (Substack). https://garymarcus.substack.com/p/ai-layoffs-productivity-and-the-klarna

Marguerit, David. 2025. "Augmenting or Automating Labor? The Effect of AI Development on New Work, Employment, and Wages." arXiv preprint arXiv:2503.19159.

Markey, Edward, and Jack Fields. 2016. The Telecommunications Act of 1996. The Hill. https://thehill.com/blogs/congress-blog/technology/272406-the-telecommunications-act-of-1996-then-now-and-beyond

Markoff, John. 1997. “Microsoft Comes to the Aid of a Struggling Apple.” The New York Times. https://archive.nytimes.com/www.nytimes.com/library/cyber/week/080797apple.html

Marshall Phelps and Cheryl Milone. 2012. LTE Standard Essential Patents Now and in the Future. Article One Partners. White paper. https://1library.net/document/y82ddr5y-lte-standard-essential-patents-now-and-in-the-future.html

Martin, Jeremy, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik Rye, and Dane Brown. 2017. "A Study of MAC Address Randomization in Mobile Devices and When it Fails." *Proceedings on Privacy Enhancing Technologies* 4: 365–383.

Martin, Matthew. 2025. "For Some Recent Graduates, the A.I. Job Apocalypse May Already Be Here." New York Times, May 30. https://www.nytimes.com/2025/05/30/technology/ai-jobs-college-graduates.html

Maskus, Keith. 2000. Intellectual Property Rights in the Global Economy. Washington, DC: Institute for International Economics.

Maslej, Nestor, et al. 2024. “The AI Index 2024 Annual Report.” AI Index Steering Committee, Institute for Human-Centered AI. Stanford University, Stanford, CA. https://hai.stanford.edu/assets/files/hai_ai-index-report-2024-smaller2.pdf

Maslej, Nestor, et al. 2025. “The AI Index 2025 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI. Stanford University, Stanford, CA. https://doi.org/10.48550/arXiv.2504.07139

Mason, Edward. 1964. Economic Concentration and the Monopoly Problem. New York: Atheneum.

Mastercard Incorporated. 2021. “2020 Annual Report.” https://s25.q4cdn.com/479285134/files/doc_financials/2020/ar/MA-2020-Annual-Report.pdf

Matsushita Electric Industrial. 1986. Matsushita Electric Industrial Co., LTD, et al. v. Zenith Radio Corp. et al., No. 83-2004. U.S. Supreme Court, March 26. https://supreme.justia.com/cases/federal/us/475/574

Mattson, Peter, et al. 2020. “MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance.” IEEE Micro 40(2): 8–16.

Maues, Julia. 2025. Banking Act of 1933 (Glass-Steagall). Federal Reserve History. https://www.federalreservehistory.org/essays/glass-steagall-act

Mavens Global. 2025. AI vs. Humans: What Jobs Will Always Require a Human Touch? https://www.linkedin.com/pulse/ai-vs-humans-what-jobs-always-require-human-touch-mavens-global-3anyc

Mazurek, Jan. 1999. Making Microchips: Policy, Globalization, and Economic Restructuring in the Semiconductor Industry. Cambridge, MA: MIT Press.

Mazzucato, Maria. 2013. The Entrepreneurial State: Debunking Public vs. Private Sector Myths. London: Anthem Press.

McAfee, Andrew 2019. More from Less New York: Scribner.

McAfee, Andrew, and Erik Brynjolfsson. 2012. “Big Data.” Harvard Business Review. https://hbr.org/2012/10/big-data-the-management-revolution

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude Elwood Shannon. 1956. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." *AI Magazine* 27(4): 12–14.

McCarty, Nolan, Keith Poole, and Howard Rosenthal. 2016. Polarized America. 2nd ed. Cambridge, MA: MIT Press.

McClain, Colleen, Kennedy, Brian, Gottfried, Jeffrey, Anderson, Monica, and Pasquini, Giancarlo. 2025. “How the U.S. Public and AI Experts View Artificial Intelligence.” Pew Research Center. https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/?itid=lk_inline_enhanced-template

McCloskey, Deirdre. 2016. Bourgeois Equality: How Ideas, Not Capital or Institutions, Enriched the World. Chicago: University of Chicago Press.

McCraw, Thomas. 1984. Prophets of Regulation. Cambridge, MA: Harvard University Press Press.

McCullough, Matthew. 2025. “16 Things to Know for Android Developers at Google I/O 2025.” Android Developers Blog. https://android-developers.googleblog.com/2025/05/16-things-to-know-for-android-developers-google-io-2025.html

McElheran, Kristina, Mu-Jeung Yang, Zachary Kroff, and Erik Brynjolfsson. 2025. "The Rise of Industrial AI in America." CES Working Paper CES-25-27. U.S. Census Bureau, Center for Economic Studies. https://www2.census.gov/library/working-papers/2025/adrm/ces/CES-WP-25-27.pdf

McKinsey & Company. 2025. “The State of AI 2025.” https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

McKinsey Global Institute. 2000. "Why the Japanese Economy is Not Growing: Micro Barriers to Productivity Growth." McKinsey Global Institute. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/asia%20pacific/why%20the%20japanese%20economy%20is%20not%20growing/mgi_why_the_japanese_economy_is_not_growing_report.pdf

McKinsey Global Institute. 2020. "The Future of Work After COVID-19." McKinsey & Company. https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-after-covid-19

McLaughlin, Kevin. 2025. Why AI Isn’t Giving Salesforce a Boost. The Information. https://www.theinformation.com/articles/ai-giving-salesforce-boost

McMahon, Liv. 2025. “ChatGPT-maker OpenAI Releases Web Browser to Rival Google.” BBC News. https://www.bbc.com/news/technology-68445981.amp

McMillan, John. 2025. “2.5D vs. 3D IC: Which Chip Packaging Tech Is Right for You?” Siemens EDA Blog. https://blogs.sw.siemens.com/semiconductor-packaging/2025/06/24/2-5d-vs-3d-ic-which-chip-packaging-technology-is-right-for-you/

McShane, Clay, and Joel Tarr. 2007. The Horse in the City: Living Machines in the Nineteenth Century. Baltimore: Johns Hopkins University Press.

Mead Corp. 2001. U.S. v. Mead Corp., No. 99-1434, U.S. Supreme Court, June 18. https://scholar.google.com/scholar_case?case=6553117666921312576

Mehta, Aashish, Jesús Felipe, and John McCombie. 2021. "Software Industry Global Growth and Productivity Dynamics." Journal of Economic Perspectives 35(2): 102-127.

Mehta, Amit. 2025. United States v. Google LLC, Remedies Opinion. U.S. District Court for the District of Columbia, September 2. https://www.courthousenews.com/wp-content/uploads/2025/09/judge-amit-mehta-google-search-monopoly-remedy-ruling.pdf

Meltzer, Allan. 2005. "Origins of the Great Inflation." Federal Reserve Bank of St. Louis Review 87(2): 145–175.

Menaldo, Victor. 2021. “Do Patents Foster International Technology Transfer? Evidence from Spanish Steelmaking, 1850–1930.” In The Battle over Patents: History and Politics of Innovation, edited by Stephen Haber and Naomi Lamoreaux. New York: Oxford University Press: 69–111.

Menaldo, Victor. 2026. Supplementary Appendix to History’s Most Revolutionary Innovation (Cambridge University Press). https://faculty.washington.edu/vmenaldo/Books/Mbook26.pdf

Menaldo, Victor and Nicolas Wittstock. 2021. “Does Technology Transfer from the US to China harm American firms and consumers? A historical and analytic investigation.” Economic and Political Studies 9(4): 417-446.

Menaldo, Victor and Nicolas Wittstock. 2025. U.S. Innovation Inequality and Trumpism. New York: Cambridge University Press.

Menand, Louis. 2023. “The Rise and Fall of Neoliberalism.” The New Yorker. https://www.newyorker.com/magazine/2023/07/24/the-rise-and-fall-of-neoliberalism

Meredith, Sam. 2025. “Meta and Google are laying a fast-growing web of mega subsea cables.” CNBC. https://www.cnbc.com/2025/07/17/meta-and-google-are-laying-a-fast-growing-web-of-mega-subsea-cables.html

Merizzi, Nicholas, Chris Thomas, and Ed Burns. 2025. The AI infrastructure reckoning: Optimizing compute strategy in the age of inference economics. Deloitte Insights, Tech Trends 2026. https://www.deloitte.com/us/en/insights/topics/technology-management/tech-trends/2026/ai-infrastructure-compute-strategy.html

Merrill, Thomas. 2002. "The Mead Doctrine." Administrative Law Review 54(3): 807–835.

Meta. 2025. “Helping Indian Brands Grow Their Cross Border Business With Meta Ad Tools.” Meta Newsroom. https://about.fb.com/news/2025/05/helping-indian-brands-grow-their-cross-border-business-with-meta-ad-tools/

Meta AI. 2024. "Introducing Llama 3: The Most Capable Openly Available LLM to Date." Meta AI Blog. https://ai.meta.com/blog/meta-llama-3/

Meta Platforms Inc. 2021. Form 10-K for the fiscal year ended December 31, 2020. SEC. https://www.sec.gov/Archives/edgar/data/1326801/000132680121000014/mti-20201231.htm

Meta Platforms Inc. 2022. Annual Report (Form 10-K). Menlo Park, CA: Meta Platforms Inc.

Meta Platforms Inc. 2023. Annual Report (Form 10-K). Menlo Park, CA: Meta Platforms Inc.

Meta Platforms Inc. 2024. "Meta Earnings Presentation Q3 2024." https://s21.q4cdn.com/399680738/files/doc_financials/2024/q3/Earnings-Presentation-Q3-2024.pdf

Meta Platforms Inc. 2025a. 2025. Q1 2025 Earnings Call Transcript/CFO outlook (CapEx guidance). https://s21.q4cdn.com/399680738/files/doc_financials/2025/q1/Transcripts/META-Q1-2025-Earnings-Call-Transcript.pdf

Meta Platforms Inc. 2025b. "Hello, Bowling Green!" Meta Data Centers. https://datacenters.atmeta.com/2025/04/hello-bowling-green/

Metz, Cade, Cecilia Kang, Sheera Frenkel, Stuart Thompson, and Nico Grant. 2024. "How Tech Giants Cut Corners to Harvest Data for AI." The New York Times. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html

Micron. 2024. DDR5 DRAM. Micron Corporation. https://www.micron.com/products/memory/dram-components/ddr5-sdram

Microsoft Corp. 2025. Form 10-K for the fiscal year ended June 30, 2024 (Intangible assets schedule). SEC. https://www.sec.gov/Archives/edgar/data/789019/000156459024015124/msft-20240630.htm

Microsoft Research. 2018. “Jabil Pilots Azure and Project Brainwave in Advanced Manufacturing Solutions.” https://www.microsoft.com/en-us/research/video/jabil-pilots-azure-project-brainwave-advanced-manufacturing-solutions/

Microsoft. 1985. Microsoft Windows Operating Environment User's Guide. Microsoft Corp. https://americanhistory.si.edu/collections/object/nmah_1695414

Microsoft. 2001. United States v. Microsoft Corp., Nos. 00-5212, 5213. US Court of Appeals, District of Columbia Circuit, June 28. https://law.justia.com/cases/federal/appellate-courts/F3/253/34

Microsoft. 2024. Azure HPC and AI Solutions. https://azure.microsoft.com/en-us/solutions/high-performance-computing/

Microsoft. 2025a “The Next Chapter of the Microsoft–OpenAI Partnership.” Official Microsoft Blog. https://blogs.microsoft.com/blog/2025/10/28/the-next-chapter-of-the-microsoft-openai-partnership

Microsoft. 2025b. “Build Apps through Conversation with Copilot.” Microsoft Learn (Power Apps documentation). https://learn.microsoft.com/en-us/power-apps/maker/canvas-apps/ai-conversations-create-app

Mih, Atah Nuh, Hung Cao, Joshua Pickard, Monica Wachowicz, and Rickey Dubay. 2023. “TransferD2: Automated Defect Detection Approach in Smart Manufacturing using Transfer Learning Techniques.” Working Paper. https://www.cs.unb.ca/~hcao3/publications/2023-coins/TransferD2.pdf

Milgrom, Paul, and John Roberts. 1990. "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities." Econometrica 58(6): 1255–1277.

Milgrom, Paul, and John Roberts. 1982. "Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis." Econometrica 50(2): 443-459.

Milgrom, Paul. 2004. Putting Auction Theory to Work. Cambridge: Cambridge University Press.

Miller, Chris. 2022. Chip War: The Fight for the World’s Most Critical Technology. New York: Scribner.

Miller, Gabby. 2024. Unpacking New NIST Guidance on Artificial Intelligence. Tech Policy Press. https://techpolicy.press/unpacking-new-nist-guidance-on-artificial-intelligence

Milli, Smitha, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca Dragan. 2025. “Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media.” PNAS Nexus 4(3): pgaf062.

Mills, Mark. 2021. The Cloud Revolution. New York: Encounter Books.

Mims, Christopher 2024. Why Every Big Tech Company Has Failed to Dethrone Nvidia as King of AI. Wall Street Journal. https://www.wsj.com/tech/ai/ai-nvidia-apple-amd-jensen-huang-software-bb581f5a

Mims, Christopher. 2025a. The New Chips Designed to Solve AI’s Energy Problem. The Wall Street Journal. https://www.wsj.com/tech/ai/the-new-chips-designed-to-solve-ais-energy-problem-1ba9cac1

Mims, Christopher. 2025b. Silicon Valley’s new strategy: Move slow and build things. The Wall Street Journal. https://www.wsj.com/tech/ai/silicon-valley-ai-infrastructure-capex-cffe0431

Miotto, Riccardo, Li Li, Brian Kidd, and Joel Dudley. 2016. "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records." Scientific Reports 6(1): 26094.

Mirowski, Philip, and Dieter Plehwe, eds. 2009. The Road from Mont Pèlerin: The Making of the Neoliberal Thought Collective. Cambridge, MA: Harvard University Press.

Misa, Thomas. 1985. "Military Needs, Commercial Realities, and the Development of the Transistor, 1948–1958." In Military Enterprise and Technological Change, edited by Merritt Roe Smith. Cambridge, MA: MIT Press, 253–287.

Mistral AI. 2024. "Mistral Large: Technical Report." Paris: Mistral AI. https://mistral.ai/

Mitchel. 1711. Mitchel v. Reynolds, 2 4 Eng. Rep. 347. Court of King's Bench. https://www.quimbee.com/cases/mitchel-v-reynolds

Mitchell, Laura, and Nicole Trotta 2025. "OFCCP To Close All Prior Section 503 and VEVRAA Compliance Reviews Following Secretary of Labor Order." Affirmative Action and OFCCP Law Advisor. https://www.affirmativeactionlawadvisor.com/2025/07/ofccp-to-close-all-prior-section-503-and-vevraa-compliance-reviews-following-secretary-of-labor-order-reviving-enforcement-activities/

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. “Model Cards for Model Reporting.” Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT): 220–229.

MLCommons. 2024. MLPerf benchmarks: Driving transparent AI performance standards. MLCommons. https://mlcommons.org/en/mlperf/

MLflow. 2024. MLflow: Open source platform for the machine learning lifecycle. https://mlflow.org/docs/latest/index.html

Mobley. 2025. Mobley v. Workday, Inc. No. 23-cv-00770-RFL. Northern District of California, June 11. https://www.govinfo.gov/content/pkg/USCOURTS-cand-3_23-cv-00770/pdf/USCOURTS-cand-3_23-cv-00770-1.pdf

Mokyr, Joel. 1990. The Lever of Riches. New York: Oxford University Press.

Mollick, Ethan. 2024. Co-Intelligence: Living and Working with AI. New York: Portfolio/Penguin.

Mollick, Ethan. n.d. One Useful Thing. Substack newsletter. https://www.oneusefulthing.org/

Montgomery, Blake and Johana Bhuiyan. 2024. “California won’t require big tech firms to test safety of AI after Newsom kills bill.” The Guardian. https://www.theguardian.com/us-news/2024/sep/29/california-governor-gavin-newsom-vetoes-ai-safety-bill

Moore, Gordon. 1965. Cramming More Components onto Integrated Circuits. Electronics 38(8): 114–117.

Mordor Intelligence. 2024. Electronic Design Automation Tools Market Analysis (2025–2030). https://www.mordorintelligence.com/industry-reports/electronic-design-automation-eda-tools-market

Moresi, Serge, and Carl Shapiro. 2012. "Any-and-All Offers in Antitrust and Consumer Protection." Antitrust 26(3): 41–45.

Motor Vehicle Manufacturers Association. 1983. *Motor Vehicle Manufacturers Association of the United States, Inc. et al. v. State Farm Mutual Automobile Insurance Co. et al.*, No. 82-354, U.S. Supreme Court, June 24. https://supreme.justia.com/cases/federal/us/463/29/

Mowery, David and Arvis Ziedonis. 2001. The Growth of Patenting and Licensing by U.S. Universities. Research Policy 30(1): 99-119.

Mowery, David, and Timothy Simcoe. 2004. "Is the Internet a U.S. Invention? An Economic and Technological History of Computer Networking." Research Policy 31(8-9): 1369-1387.

Mowery, David, Richard Nelson, Bhaven Sampat, and Arvids Ziedonis. 2001. “The Growth of Patenting and Licensing by U.S. Universities: An Assessment of the Effects of the Bayh–Dole Act of 1980.” Research Policy 30(1): 99–119.

Mowery, David, and Bhaven Sampat. 2005. “The Bayh–Dole Act of 1980 and University–Industry Technology Transfer.” Journal of Technology Transfer 30(1–2): 115–127.

Mowery, David, Richard Nelson, Bhaven Sampat, and Arvids Ziedonis. 2004. Ivory Tower and Industrial Innovation. Stanford: Stanford University Press.

Mozilla. 2019. "Firefox Now Available with Enhanced Tracking Protection by Default." Mozilla Blog. https://blog.mozilla.org/en/products/firefox/firefox-now-available-with-enhanced-tracking-protection-by-default/

Mozingo, Tom. 2020. Revisiting the Enforceability of Online Contracts: The Need for Unambiguous Assent to Inconspicuous Terms. Seattle University Law Review 43(3): 1065.

Mukesh, Sagarika, and Jingyun Zhang. 2022. "A Review of the Gate-All-Around Nanosheet FET Process Opportunities." Electronics 11(21): 3589.

Muris, Timothy. 2023. "Neo-Brandeisian Antitrust: Repeating History's Mistakes." *Regulation* 46(1): 6–9.

Musk, Elon. 2023. “To address extreme levels of data scraping and system manipulation, we’ve applied the following temporary limits.” X (formerly Twitter). https://x.com/elonmusk/status/1675187969420828672%5C%3E%5B2

Myung, Hyung, Junsung Lim, and David Goodman. 2006. "Single Carrier FDMA for Uplink Wireless Transmission." IEEE Vehicular Technology Magazine 1(3): 30-38.

Na, Hang Ryeol 2025. “Environmental Governance Under Trump: From EPA Decline to Judicial Ascendancy.” NiCHE. https://niche-canada.org/2025/06/25/environmental-governance-under-trump-from-epa-decline-to-judicial-ascendancy/

Nada Golmie, Alhussein Abouzeid, Thyagarajan Nandagopal, Murat Torlak, Marc Leh, and Miller Higgins. 2023. NextG Communications Research and Development Gaps Report (NIST

Special Publication 1293). NIST. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1293.pdf

Nagabandi, Aravind, Kurt Konolige, Sergey Levine, and Vijay Kumar. 2020. "Deep Dynamics Models for Learning Dexterous Manipulation." *Proceedings of the Conference on Robot Learning*, PMLR 100: 1101–1112.

Nagle, Frank, Robert Seamans, and Steven Tadelis. 2020. Transaction Cost Economics in the Digital Economy: A Research Agenda. Harvard Business School Strategy Unit Working Paper No. 21-009.

Nagler, Markus, Monika Schnitzer, and Martin Watzinger. 2022. "Fostering the Diffusion of General Purpose Technologies: Evidence from the Licensing of the Transistor Patents." *Journal of Industrial Economics* 70(4): 828-875.

NAHB 2025. “DOE Suspends Energy Efficiency Mandates on Key Home Appliances.” National Association of Home Builders. https://www.nahb.org/blog/2025/02/doe-suspends-energy-efficiency-mandates-on-key-home-appliances

Nahra, Kirk, Jennings, Molly, Straky, Andrew, Jessani, Ali. 2025. Year in Review: 2024 BIPA Litigation Takeaways. WilmerHale. https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20250219-year-in-review-2024-bipa-litigation-takeaways

Naidu, Richa, Martin Coulter and Jason Lange. 2023. "Focus: ChatGPT Fever Spreads to U.S. Workplace, Sounding Alarm for Some." Reuters. https://www.reuters.com/technology/chatgpt-fever-spreads-us-workplace-sounding-alarm-some-2023-08-11/

Nanda, Arun. 2025. Azure OpenAI. Datacamp. https://www.datacamp.com/tutorial/azure-openai-getting-started

NARA. 2024. Code of Federal Regulations, Title 37: Patents, Trademarks, and Copyrights. Office of the Federal Register. https://www.govinfo.gov/content/pkg/CFR-2024-title37-vol1/pdf/CFR-2024-title37-vol1.pdf

Narayanan, Arvind, and Sayash Kapoor. 2025. “AI as Normal Technology.” Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/ai-as-normal-technology

NCCUSL. 1999. Uniform Computer Information Transactions Act. Chicago: NCCUSL.

National Science Board. 2020. *The State of U.S. Science and Engineering 2020*. Alexandria, VA: National Science Foundation. https://ncses.nsf.gov/pubs/nsb20201

National Science Board. 2022. *The State of U.S. Science and Engineering 2022*. Alexandria, VA: National Science Foundation. https://ncses.nsf.gov/pubs/nsb20221

NaturalGas.org. 2013. The History of Regulation. http://naturalgas.org/regulation/history/

Nayak, Pandu. 2019. "Understanding searches better than ever before." The Keyword (Google). https://blog.google/products/search/search-language-understanding-bert/

NCAA. 1984. NCAA v. Board of Regents of the University of Oklahoma, 468 U.S. 85. https://supreme.justia.com/cases/federal/us/468/85

NCES. 2000. Internet Access in U.S. Public Schools and Classrooms: 1994–99. Stats in Brief. https://nces.ed.gov/pubs2000/2000086.pdf

NCRALP. 1979. Report to the President and the Attorney General of the National Commission for the Review of Antitrust Laws and Procedures. Washington, DC: U.S. Government Printing Office.

Nebuly. 2024. OpenAI GPT-4 API pricing: The drops from 2023 to 2024. https://www.nebuly.com/blog/openai-gpt-4-api-pricing

Nellis, Stephen. 2023. “Qualcomm to supply Apple with 5G chips until 2026 under new deal.” Reuters. https://www.reuters.com/technology/qualcomm-supply-apple-with-5g-chips-until-2026-under-new-deal-2023-09-11/

Nellis. Stephen. 2025. “Stargate UAE AI datacenter to begin operation in 2026.” Reuters. https://www.reuters.com/business/media-telecom/stargate-uae-ai-datacenter-begin-operation-2026-2025-05-22/

Nelson, Richard, and Edmund Phelps. 1966. "Investment in Humans, Technological Diffusion, and Economic Growth." American Economic Review 56(1/2): 69-75.

Nelson, Richard. 1959. "The Simple Economics of Basic Scientific Research." Journal of Political Economy 67(3): 297–306.

Neumark, David, and William Wascher. 2008. Minimum Wages. Cambridge, MA: MIT Press.

Nicholas, Gabriel. 2022. “Shedding Light on Shadowbanning.” Center for Democracy and Technology. https://cdt.org/insights/shedding-light-on-shadowbanning/.

Nickelsburg, Monica, Joshua McNichols, Lucy Soucek, Alec Cowan, and Carol Smith. 2025. Is our AI obsession good for small town America? Booming Podcast. https://www.kuow.org/stories/is-our-ai-obsession-good-for-small-town-america

Nielsen, Rasmus Kleis, and Sarah Anne Ganter. 2022. The Power of Platforms. New York: Oxford University Press.

Nikolić, Igor. 2017. “Who needs injunctions? Alternative remedies in standard essential patents disputes.” Journal of Intellectual Property Law and Practice 12(2): 126–135.

Niskanen, William. 1988. Reaganomics. New York: Oxford University Press.

NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

NIST. 2024. Interagency Committee on Standards Policy (ICSP). https://www.nist.gov/standardsgov/interagency-committee-standards-policy-icsp

Nixon, Richard. 1971. Executive Order 11615: Stabilization of Prices, Rents, Wages, and Salaries. Federal Register 36(159): 15727.

Nordhaus, William D. 2004. "Schumpeterian Profits in the American Economy: Theory and Measurement." NBER Working Paper No. 10433. Cambridge, MA: National Bureau of Economic Research.

Nordhaus, William. 2007. "Two Centuries of Productivity Growth in Computing." The Journal of Economic History 67(1): 128–159.

Northern Pacific Railway. 1958. Northern Pacific Railway Co. v. United States, 356 U.S. 1. https://supreme.justia.com/cases/federal/us/356/1

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." Science 3816654: 187–192.

NSB. 2020. Science and Engineering Indicators 2020. "Academic Research and Development." Alexandria, VA: National Science Foundation.

NSB. 2024. Science and Technology: Public Perceptions and Awareness in the United States. NSF. https://ncses.nsf.gov/pubs/nsb20244/assets/nsb20244.pdf

Nuechterlein, Jonathan, and Philip Weiser. 2005. Digital Crossroads. Cambridge, MA: MIT Press.

Nuvolari, Alessandro. 2009. "Understanding Successive Industrial Revolutions: A 'Collective Invention' Perspective." *Rivista di Politica Economica* 99(3-4): 9–40.

Nuñez, Michael. 2025. "Anthropic Launches Enterprise 'Agent Skills' and Opens the Standard, Challenging OpenAI in Workplace AI." VentureBeat. https://venturebeat.com/ai/anthropic-launches-enterprise-agent-skills-and-opens-the-standard

NVIDIA Developer. 2025. Deep Learning Frameworks. NVIDIA Developer. https://developer.nvidia.com/deep-learning-frameworks

NVIDIA. 2024a. “NVIDIA H100 Tensor Core GPU Architecture White Paper.” https://www.advancedclustering.com/wp-content/uploads/2022/03/gtc22-whitepaper-hopper.pdf

NVIDIA. 2024b. “NVIDIA Blackwell Platform Arrives to Power a New Era of Computing.” NVIDIA Investor Relations Press Release. https://investor.nvidia.com/news/press-release-details/2024/NVIDIA-Blackwell-Platform-Arrives-to-Power-a-New-Era-of-Computing/default.aspx

NVIDIA. 2025a. “CFO Commentary on Fourth Quarter and Fiscal 2025 Results.” Nvidia Investor Relations. https://investor.nvidia.com/files/doc_financials/2025/Q425/Q4FY25-CFO-Commentary.pdf

NVIDIA. 2024c. “NVIDIA Ethernet Networking Accelerates World’s Largest AI Supercomputer, Built by xAI.” NVIDIA Newsroom. https://nvidianews.nvidia.com/news/spectrum-x-ethernet-networking-xai-colossus

NVIDIA. 2025b. NVIDIA TensorRT. Nvidia Documentation Hub. https://docs.nvidia.com/tensorrt/index.html

NVIDIA. 2025c. “NVIDIA Announces DGX Cloud Lepton to Connect Developers to NVIDIA’s Global Compute Ecosystem.” Nvidia Press Release. https://nvidianews.nvidia.com/news/nvidia-announces-dgx-cloud-lepton-to-connect-developers-to-nvidias-global-compute-ecosystem

NVM Express. 2024. "About NVMe Technology." NVM Express Inc. https://nvmexpress.org/specifications/

Nye, Joseph. 2004. Power in the Global Information Age: From Realism to Globalization. New York: Routledge.

Nylen, Leah. 2023. “Google Pays Apple 36% of Search Revenue from Safari, Witness Says.” Bloomberg. https://www.bloomberg.com/news/articles/2023-11-13/apple-gets-36-of-google-revenue-from-search-deal-witness-says

Nynex. 1998. Nynex Corp. v. Discon, Inc., No. 96-1570. US Supreme Court, December 14. https://supreme.justia.com/cases/federal/us/525/128

O'Hara, Kieron, and Wendy Hall. 2018. Four Internets: The Geopolitics of Digital Governance. CIGI Paper No. 206. Centre for International Governance Innovation.

O'Reilly, Tim. 2021. "Data Is the New Sand." The Information. https://www.theinformation.com/articles/data-is-the-new-sand.

O'Rourke, Kevin, and Jeffrey Williamson. 1999. Globalization and History: The Evolution of a Nineteenth-Century Atlantic Economy. Cambridge, MA: MIT Press.

Obama, Barack. 2011. “Executive Order 13563—Improving Regulation and Regulatory Review." 76 FR: 3821. https://www.federalregister.gov/documents/2011/01/21/2011-1385/improving-regulation-and-regulatory-review

OECD Statistics. 2025. “R&D Spending Growth Slows in OECD, Surges in China: Government Support for Energy and Defence R&D Rises Sharply.” OECD, March 30.

OECD. 2023a. Trade in Value Added (TiVA) Indicators: United States. OECD. https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/trade-in-value-added/tiva-2023-USA.pdf

OECD. 2023b. Trade in Value Added (TiVA) Database, Version 2023. OECD. https://www.oecd.org/sti/ind/inter-country-input-output-tables.htm.

OECD. 2024. OECD Principles on Artificial Intelligence. https://oecd.ai/en/ai-principles

Ohlhausen, Maureen. 2013. "Antitrust Enforcement in High Technology Markets," FTC Speech, October 16. CCIA Annual Washington Caucus. https://www.ftc.gov/news-events/news/speeches/antitrust-enforcement-high-technology-markets

OIRA. 2024. "Historical Reports: OIRA Concluded Reviews (1981–2023)." Washington, D.C.: General Services Administration.

Oladayo, Oluwatosin. 2024. "Future-proofing AI storage infrastructure: Managing scale, performance and data diversity." Open Access Research Journal of Science and Technology 12(1): 170-185.

Ollinger, Michael, and Valerie Mueller. 2003. Managing for Safer Food. Agricultural Economic Report No. 817. Department of Agriculture, Economic Research Service. https://ers.usda.gov/sites/default/files/_laserfiche/publications/41496/18901_aer817.pdf?v=26386

OMB. 1996. Economic Analysis of Federal Regulations under Executive Order 12866. OMB Memorandum for Members of the Regulatory Working Group. https://trumpwhitehouse.archives.gov/wp-content/uploads/2017/11/1996-Memorandum-for-Members-of-the-Regulatory-Working-Group-Economic-Analysis-of-Federal-Regulations-Under-Executive-Order-12866.pdf

OMB. 1998. "Circular A-119: Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities." OMB. FR 63: 8546–55.

https://www.federalregister.gov/documents/1998/02/19/98-4177/omb-circular-a-119-federal-participation-in-the-development-and-use-of-voluntary-consensus-standards

OMB. 2003. "Circular A-4: Regulatory Analysis." OMB. FR 68: 58366. https://www.federalregister.gov/documents/2003/10/09/03-25606/circular-a-4-regulatory-analysis

OPC Foundation. 2023. OPC UA Interoperability for Industrie 4.0 and IoT. OPC Foundation. https://opcfoundation.org/wp-content/uploads/2023/05/OPC-UA-Interoperability-For-Industrie4-and-IoT-EN.pdf

OpenAI. 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774

OpenAI. 2025a. “Introducing Operator.” OpenAI. https://openai.com/index/introducing-operator/

OpenAI. 2025b. “Computer-Using Agent.” OpenAI. https://openai.com/index/computer-using-agent/

OpenAI. 2025c. Introducing GPT-5. OpenAI. https://openai.com/en-GB/index/introducing-gpt-5

OpenAI. 2025d. “Update to GPT-5 System Card: GPT-5.2.” OpenAI Research. https://openai.com/index/gpt-5-system-card-update-gpt-5-2/

OpenAI. 2025e. "OpenAI Co-Founds the Agentic AI Foundation under the Linux Foundation." OpenAI Blog, December. https://openai.com/index/agentic-ai-foundation/

OpenAI. 2025f. “Pricing.” OpenAI API Documentation. https://platform.openai.com/docs/pricing

OpenAI. 2025g. “Stargate advances with 4.5 GW partnership with Oracle.” https://openai.com/index/stargate-advances-with-partnership-with-oracle/

OpenAI. n.d. Pricing and Business Model. https://chatgpt.com/pricing

Ord, Toby. 2025. “The Scaling Paradox.” https://www.tobyord.com/writing/the-scaling-paradox

Oregon Steam Navigation. 1873. Oregon Steam Navigation Company v. Winsor, 87 U.S. 64, U.S. Supreme Court. https://supreme.justia.com/cases/federal/us/87/64/

Orgvue. 2025. “55% of Businesses Admit Wrong Decisions in Making Employees Redundant When Bringing AI into the Workforce.” Orgvue.com. https://www.orgvue.com/news/55-of-businesses-admit-wrong-decisions-in-making-employees-redundant-when-bringing-ai-into-the-workforce/

Osawa, Juro and Qianer Liu. 2025. “DeepSeek: a National Treasure in China, Now Closely Guarded.” The Information. https://www.theinformation.com/articles/deepseek-national-treasure-china-now-closely-guarded?rc=tuorsb.

OSHA. 2012. Hazard Communication Standard: Safety Data Sheets. DoL. https://www.osha.gov/hazcom

Ouyang, Long, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35:27730–27744.

Owens, Kellie, Zachary Griffen, and Lasya Damaraju, 2025. "Managing a 'Responsibility Vacuum' in AI Monitoring and Governance in Healthcare: A Qualitative Study." BMC Health Services Research 25(1): 1217.

Ozkan, Serdar, and Nicholas Sullivan. 2025. “Is AI Contributing to Rising Unemployment? Evidence from Occupational Variation.” Federal Reserve Bank of St. Louis—On the Economy Blog. https://www.stlouisfed.org/on-the-economy/2025/aug/is-ai-contributing-unemployment-evidence-occupational-variation.

O’Mara, Margaret 2019. The Code. New York: Penguin Press.

P1 Security. 2025. 4G LTE Architecture and Security Explained: Protocols, Attack Vectors, and Persistent Vulnerabilities. P1 Security Blog. https://www.p1sec.com/blog/4g-architecture-and-security

Pacific Bell. 2009. Pacific Bell Telephone Co. v. linkLine Communications, Inc., No. 07-512. US Supreme Court, February 25. https://supreme.justia.com/cases/federal/us/555/438

Page, Larry. 2014. "Lenovo to Acquire Motorola Mobility." Google Blog. https://blog.google/inside-google/company-announcements/lenovo-to-acquire-motorola-mobility/

Page, William, and John Lopatka. 2007. The Microsoft Case: Antitrust, High Technology, and Consumer Welfare. Chicago: University of Chicago Press.

Palazzolo, Stephanie, and Cory Weinberg. 2025. “The Little-Known Startup That Has Surged Past Scale AI—Without Any Investors.” The Information. https://www.theinformation.com/articles/little-known-startup-surged-past-scale-ai-without-investors

Palazzolo, Stephanie, and Sri Muppidi. 2025. "How Anthropic and OpenAI Are Developing AI 'Co-Workers'." The Information. https://www.theinformation.com/articles/anthropic-openai-developing-ai-co-workers

Palmer, Annie. 2025. "AI Will Shrink Amazon's Workforce in the Coming Years, CEO Jassy Says." CNBC. https://www.cnbc.com/2025/06/17/ai-amazon-workforce-jassy.html

Palmer, Maija. 2019. “Does Europe need to fix ‘monkey money’ stock options?” Sifted. https://sifted.eu/articles/europe-fix-stock-options-notoptional-employee-ownership

Parbel, Matthias. 2025. Snowflake and Salesforce launch open standard for standardized AI data (Open Semantic Interchange). Heise Online. https://www.heise.de/en/news/Snowflake-and-Salesforce-launch-open-standard-for-standardised-AI-data-10669447.html

Pardau, Stuart. 2018. “The California Consumer Privacy Act: Towards a European-Style Privacy Regime in the United States?” Journal of Technology Law and Policy 23(1): 68-114.

Parker 1978. Parker, Acting Commissioner of Patents and Trademarks v. Flook, No. 77-642, U.S. Supreme Court, June 22. https://scholar.google.com/scholar_case?case=12542933152070461616

Parker, Edward, Benjamin Miller, and Colin Levaunt. 2025. Who Could Fund Future Artificial Intelligence Development? Expert Insights. RAND Corporation. https://www.rand.org/pubs/perspectives/PEA3701-1.html

Parker, Geoffrey, and Marshall Van Alstyne. 2005. "Two-Sided Network Effects: A Theory of Information Product Design." Management Science 51(10): 1494-1504.

Parker, Geoffrey, Marshall Van Alstyne, and Sangeet Paul Choudary. 2016. Platform Revolution. New York: W.W. Norton.

Partovi, Hadi. 2025. No, AI isn’t killing computer science. It’s making it essential. The Information. https://www.theinformation.com/articles/ai-killing-computer-science-making-essential

Patel, Dwarkesh. 2025. “Satya Nadella – Microsoft’s AGI Plan and Quantum Breakthrough.” The Dwarkesh Podcast. https://www.dwarkesh.com/p/satya-nadella

Patel, Marc Sorel, and Pankaj Sachdeva. 2025. What Is a Data Center? McKinsey Explainer. McKinsey & Company. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-a-data-center#/

PatentTrack. 2024. Qualcomm patent portfolio overview. https://patentrack.com/strategic-patent-integration-qualcomms-portfolio/

Patterson, David and John Hennessy. 2021. Computer Organization and Design: The Hardware/Software Interface (6th ed.). San Francisco: Morgan Kaufmann.

Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. “Carbon Emissions and Large Neural Network Training.” arXiv:2104.10350. https://arxiv.org/pdf/2104.10350.pdf

Paxson, David. 1981. “Potential Impact of Motor Carrier Act of 1980 on Railroad Industry.” Transportation Research Record 804: 34–39.

PCPIP. 1883. Paris Convention for the Protection of Industrial Property of March 20, 1883, as revised at Brussels on December 14, 1900, at Washington on June 2, 1911, at The Hague on November 6, 1925, at London on June 2, 1934, at Lisbon on October 31, 1958, and at Stockholm on July 14, 1967. WIPO. https://www.wipo.int/wipolex/en/text/288514

Peck, Emily. 2025. "AI and College Grads: The Job Crunch Is Real." Axios. https://www.axios.com/2025/05/29/ai-college-grads-work-jobs

Peltzman, Sam. 1976. "Toward a More General Theory of Regulation." Journal of Law and Economics 19(2): 211-240.

Perez, Carlota. 2002. Technological Revolutions and Financial Capital. Cheltenham, UK: Edward Elgar Publishing.

Peteraf, Margaret. 1993. "The Cornerstones of Competitive Advantage: A Resource-Based View." Strategic Management Journal 14(3): 179-191.

Peterson, Kyle. 2011. “Boeing Dreamliner Makes First Delivery after Three-Year Delay.” Reuters. https://www.reuters.com/article/2011/01/20/us-boeing-dreamliner-idUSTRE70J2UX20110120/

Petzold, Charles. 2022. Code: The hidden language of computer hardware and software (2nd ed.). Redmond, WA: Microsoft Press.

Philadelphia National Bank. 1963. United States v. Philadelphia National Bank, No. 83. US Supreme Court, June 17. https://supreme.justia.com/cases/federal/us/374/321/

Pigou, Arthur. 1920. The Economics of Welfare. London: Macmillan.

Pindyck, Robert. 2004. "Mandatory Unbundling and Irreversible Investment in Telecom Networks." Review of Network Economics 3(3): 274-298.

Pindyck, Robert. 2007. “Mandatory Unbundling and Irreversible Investment in Telecom Networks.” Review of Network Economics 6(3): 274–298.

Pineda, Michael. 2021. “Major Suppliers of Apple: Inside Its Supply Chain.” Profolus. https://www.profolus.com/topics/major-suppliers-of-apple-inside-its-supply-chain/

Piosenka, Gerald Vincent, and Ronald Chandos. 1991. Unforgeable Personal Identification System. (US Patent No. 4,993,068). USPTO. https://patents.google.com/patent/US4993068A/en

Pitofsky, Robert 2001. The Essential Facilities Doctrine Under U.S. Antitrust Law. Antitrust Law Journal 70(2): 443-462.

PM Insights. 2025. “Anthropic Approaches $7B Run Rate in 2025, Outpaces OpenAI.” PM Insights Weekly Edition and Stock Spotlight. https://www.pminsights.com/insights/anthropic-approaches-7b-run-rate-in-2025-outpaces-openai

Polaroid 1986. Polaroid Corp. v. Eastman Kodak Co., No. 76-1634-Z, U.S. District Court, D. Massachusetts, October 11. https://law.justia.com/cases/federal/district-courts/FSupp/641/828/1482978

Pooley, James. 2015. Secrets. Palo Alto, CA: Verus Press.

Porter, Michael. 1980. *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press.

Porter, Michael. 1985. Competitive Advantage: Creating and Sustaining Superior Performance. New York: Free Press.

Porter, Michael, and James Heppelmann. 2014. “How Smart, Connected Products Are Transforming Competition.” Harvard Business Review. https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition

Porter, Michael, and Mariko Sakakibara. 2004. "Competition in Japan." Journal of Economic Perspectives 18(1): 27–50.

Posner, Eric, and E. Glen Weyl. 2018. Radical Markets. Princeton, NJ: Princeton University Press.

Posner, Richard. 1976. Antitrust Law: An Economic Perspective. Chicago: University of Chicago Press.

Posner, Richard A. 1979. The Chicago School of Antitrust Analysis, 127 University of Pennsylvania Law Review 925.

Posner, Richard. 2001. Antitrust Law. 2nd ed. Chicago: University of Chicago Press.

Prabhakar, Raghu et al. 2024. “SambaNova SN40L: Scaling the AI Memory Wall with Dataflow and Composition of Experts.” arXiv preprint arXiv:2405.12345. https://arxiv.org/abs/2405.12345

Pregelj, Vladimir. 2001. Most-Favored-Nation Status of the People's Republic of China. CRS Report RL30225. https://digital.library.unt.edu/ark:/67531/metacrs2000/m2/1/high_res_d/RL30225_2001Jun07.pdf

Procore Technologies 2024. "Procore Launches Procore AI with New Agents to Boost Construction Management Efficiency." Press release. https://www.procore.com/press/procore-launches-procore-ai-with-new-agents-to-boost-construction-management-efficiency

Procore Technologies. n.d. "Procore Insights." https://www.procore.com/insights

Provost, Foster, and Tom Fawcett. 2013. Data Science for Business. Sebastopol, California: O'Reilly Media.

PwC. 2025. "PwC's 2025 Digital Trends in Operations Survey." PwC. https://www.pwc.com/us/en/services/consulting/business-transformation/digital-supply-chain-survey.html

Pérez-Cruz, Fernando, Prenio, Jermy, Restoy, Fernando, and Yong, Jeffery. 2025. Managing explanations. FSI Occasional Paper No. 24. Bank for International Settlements. https://www.bis.org/fsi/fsipapers24.pdf

Qi, Yuan, Thomas Minka, Rosalind Picard, and Zoubin Ghahramani. 2004. Predictive Automatic Relevance Determination by Expectation Propagation. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*. Banff, Canada, 846–853.

Qualcomm. 2019. FTC v. Qualcomm Inc., 414 F. Supp. 3d 1101 (N.D. Cal.). https://law.justia.com/cases/federal/district-courts/california/candce/3:2017cv01225/284890/218/

Qualcomm. 2020. FTC v. Qualcomm, Inc., No. 19-16122, 9th Cir., Aug. 11. https://law.justia.com/cases/federal/appellate-courts/ca9/19-16122/19-16122-2020-08-11.html

Qualcomm Inc. 2006. "Annual Report on Form 10-K for the Fiscal Year Ended September 24, 2006." SEC. https://www.sec.gov/Archives/edgar/data/804328/000093639206000996/a24679e10vk.htm

Qualcomm Inc. 2008. "LTE/WiMax Patent Licensing Statement." Qualcomm Public Policy Documents. https://www.qualcomm.com/media/documents/files/lte-wimax-patent-licensing-statement.pdf

Qualcomm Inc. 2014. Annual Report (Form 10-K). San Diego, CA: Qualcomm Inc. SEC. https://www.sec.gov/Archives/edgar/data/804328/000123445214000320/qcom10-k2014.htm

Qualcomm Inc. 2017. "Qualcomm 5G NR Royalty Terms Statement." Qualcomm Public Releases, November 19. https://www.qualcomm.com/media/documents/qualcomm-5g-nr-royalty-terms-statement

Qualcomm Inc. 2018. Annual Report (Form 10-K). San Diego, CA: Qualcomm Inc. SEC. https://www.sec.gov/Archives/edgar/data/804328/000172894918000095/qcom10-k2018.htm

Qualcomm Inc. 2019a. FTC v. Qualcomm Inc., Case No. 17-CV-00220-LHK (Findings of Fact and Conclusions of Law). U.S. District Court, Northern District of California, May 21. https://www.ftc.gov/system/files/documents/cases/qualcomm_findings_of_fact_and_conclusions_of_law.pdf

Qualcomm Inc. 2019b. "Qualcomm and Apple agree to drop all litigation." Press release. https://www.qualcomm.com/news/releases/2019/04/qualcomm-and-apple-agree-drop-all-litigation

Qualcomm Inc. 2019c. "Qualcomm Earnings Release Available on Company's Investor Relations Website." Press release, May 1. https://investor.qualcomm.com/news-events/press-releases/news-

details/2019/Qualcomm-Earnings-Release-Available-on-Companys-Investor-Relations-Website-05-01-2019/default.aspx

Qualcomm Inc. 2020. Annual Report (Form 10-K). San Diego, CA: Qualcomm Inc. SEC. https://www.sec.gov/Archives/edgar/data/804328/000172894920000067/qcom-20200927.htm

Qualcomm Inc. 2022a. "Qualcomm Incorporated's Response to the Competition and Market Authority's Public Consultation on Retaining the EU Horizontal Block Exemption Regulations." https://assets.publishing.service.gov.uk/media/625009d98fa8f54a841bbe09/Qualcomm.pdf

Qualcomm Inc. 2022b. "Annual Report on Form 10-K for the Fiscal Year Ended September 25, 2022." SEC. https://www.sec.gov/Archives/edgar/data/804328/000080432822000021/qcom-20220925.htm

Qualcomm Inc. 2023. Annual Report (Form 10-K). San Diego, CA: Qualcomm Inc. SEC. https://www.sec.gov/Archives/edgar/data/804328/000080432824000017/qcom-20231224.htm

Qualcomm Inc. 2024. Annual Report (Form 10-K). San Diego, CA: Qualcomm Inc. SEC. https://www.sec.gov/Archives/edgar/data/804328/000080432824000075/qcom-20240929.htm

Quinn, Gene. 2014. "Freeman-Walter-Abele: A Tortured History of Software Eligibility." IPWatchdog 2 https://ipwatchdog.com/2014/12/02/freeman-walter-abele-a-tortured-history-of-software-eligibility/id=52271/

Raab, Ben. 2025. "The $1,999 Liberty Phone Is Made in America. Its Creator Explains How." Wall Street Journal. https://www.wsj.com/tech/personal-tech/liberty-phone-purism-made-in-america-b4074c89

Radauskas, Gintaras. 2025. Tech jobs in 2025: it's not looking good, and AI will make it even worse. Cybernews. https://cybernews.com/editorial/tech-jobs-in-2025-ai-upskilling/

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language Models Are Unsupervised Multitask Learners.” OpenAI Blog. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raff, Daniel, and Lawrence Summers. 1987. "Did Henry Ford Pay Efficiency Wages?" Journal of Labor Economics 5(4): S57-S86.

Raitano, Lucy, and Amanda Cooper. 2025. "Five Debt Hotspots in the AI Data Centre Boom." Reuters. Reprinted in U.S. News and World Report. https://money.usnews.com/investing/news/articles/2025-11-05/five-debt-hotspots-in-the-ai-data-centre-boom

Raman, Aneesh. 2025. “I’m a LinkedIn Executive. I See the Bottom Rung of the Career Ladder Breaking.” New York Times. https://www.nytimes.com/2025/05/19/opinion/linkedin-ai-entry-level-jobs.html

Ramaswamy, Sridhar, and Paul Muret. 2018. “Introducing Google Marketing Platform.” Google Marketing Platform Blog. https://blog.google/products/marketingplatform/360/introducing-google-marketing-platform

Rambus 2008. Rambus Inc. v. Federal Trade Commission. No. 07-1086. US Court of Appeals, D.C. Circuit, April 22. https://law.justia.com/cases/federal/appellate-courts/cadc/07-1086/07-1086-1112217-2011-03-24.html

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8821-8831.

Ramkumar, Sudha. 2025. "ISO and ASTM Standards: An Overview." Microbe Investigations Switzerland. https://microbe-investigations.com/understanding-the-differences-between-iso-and-astm-standards/

Rana, Anurag and Andrew Girard. 2024. Big tech 2025 capex may hit $200 billion as gen-AI demand booms. Bloomberg Intelligence. https://www.bloomberg.com/professional/insights/technology/big-tech-2025-capex-may-hit-200-billion-as-gen-ai-demand-booms/

Rao, Rohit, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. 2021. MSA Transformer. Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8844-8856.

Ratner, Alexander, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. The VLDB Journal, 29(2): 709-730.

Reagan, Ronald. 1981a. Executive Order 12291: Federal Regulation. 46 FR: 13193. https://www.federalregister.gov/executive-order/12291

Reagan, Ronald. 1981b. "Statement on U.S. Trade Policy." Public Papers of the Presidents of the United States: Ronald Reagan (614–16). Washington, D.C.: US Government Printing Office.

Reagan, Ronald. 1982. Radio Address to the Nation on International Free Trade. Ronald Reagan Presidential Library. https://www.presidency.ucsb.edu/documents/radio-address-the-nation-international-free-trade

Reagan, Ronald. 1986. Radio Address to the Nation on Free and Fair Trade. Ronald Reagan Presidential Library. https://www.reaganlibrary.gov/archives/speech/radio-address-nation-free-and-fair-trade-1

Recon Analytics. 2018. How America's 4G leadership propelled the U.S. economy. Recon Analytics. https://api.ctia.org/wp-content/uploads/2018/04/Recon-Analytics_How-Americas-4G-Leadership-Propelled-US-Economy_2018.pdf

Register.com 2004. Register.com, Inc. v. Verio, Inc. 2004. No. 00-9596. U.S. Court of Appeals, Second Circuit, January 23. https://law.justia.com/cases/federal/appellate-courts/F3/356/393/576556/

Reiter. 1979. Reiter v. Sonotone Corp., 442 U.S. 330. U.S. Supreme Court, June 18. https://supreme.justia.com/cases/federal/us/442/330

Research and Markets. 2025. Automated Optical Inspection (AOI) Systems – Global Strategic Business Report. https://www.researchandmarkets.com/reports/4804449/automated-optical-inspection-aoi-systems

Reshoring Initiative. 2023. "2022 Reshoring and FDI Data Report." Reshoring Initiative. https://reshorenow.org/content/pdf/2022_Data_Report.pdf

Reuters. 2023. "EXPLAINER—What does Twitter 'rate limit exceeded' mean for users?" Reuters. https://finance.yahoo.com/news/explainer-does-twitter-rate-limit-145553424.html

Reuters. 2025a. "Elon Musk's xAI buys new property in Memphis amid supercomputer expansion." Reuters. https://www.reuters.com/technology/artificial-intelligence/elon-musks-xai-buys-new-property-memphis-amid-supercomputer-expansion-2025-03-07/

Reuters. 2025b. "'Neocloud' Crusoe secures $11.6 billion funding for Texas data center. Reuters. https://www.reuters.com/technology/openais-biggest-data-center-secures-116-billion-funding-wsj-reports-2025-05-21/

Ricardo, David. 1817. On the Principles of Political Economy and Taxation. London: John Murray.

Richards, Porter. 2025. The E-SIGN Act in the Age of Smart Contracts and AI: Challenges and Opportunities. University of Cincinnati Intellectual Property and Computer Law Journal. https://ucipclj.org/2025/04/16/the-e-sign-act-in-the-age-of-smart-contracts-and-ai-challenges-and-opportunities/

Richman, Sheldon. 1988. The Reagan Record on Trade: Rhetoric vs. Reality. Cato Institute Policy Analysis No. 107. https://www.cato.org/sites/cato.org/files/pubs/pdf/pa107.pdf

Richter, Felix. 2023. “Chart: The Trillion-Dollar App Economy.” Statista. https://www.statista.com/chart/30364/sales-and-billings-facilitated-by-the-app-store/

Riordan, Michael, and Lillian Hoddeson. 1997. Crystal Fire. New York: W. W. Norton and Company.

Rives, Alex, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118(15): e2016239118.

Robotics Tomorrow. 2025. Unitree Showcases Advanced High-Mobility Robots at CES 2025. https://www.roboticstomorrow.com/news/2025/01/08/unitree-showcases-advanced-high-mobility-robots-at-ces-2025/23818

Rochet, Jean-Charles, and Jean Tirole. 2003. "Platform Competition in Two-Sided Markets." Journal of the European Economic Association 1(4): 990–1029.

Rochet, Jean-Charles, and Jean Tirole. 2006. Two-Sided Markets: A Progress Report. RAND Journal of Economics 37(3): 645–667.

Rockoff, Hugh. 1984. Drastic Measures. Cambridge: Cambridge University Press.

Rodrik, Dani. 2004. “Industrial Policy for the Twenty-First Century.” CEPR Discussion Paper No. 4767. https://cepr.org/publications/dp4767.

Rodrik, Dani. 2006. "Goodbye Washington Consensus, Hello Washington Confusion?" Journal of Economic Literature 44(4): 973–987.

Rodrik, Dani. 2011. The Globalization Paradox: Democracy and the Future of the World Economy. New York: W. W. Norton.

Roessner, David, Jennifer Bond, Sumiye Okubo, and Mark Planting. 2013. "The economic impact of licensed commercialized inventions originating in university research." Research Policy 42(1): 23–34.

Rogers, Everett. 1962. Diffusion of Innovations. New York: Free Press of Glencoe.

Rogoff, Kenneth. 2025. Our Dollar, Your Problem: An Insider's View of Seven Turbulent Decades of Global Finance, and the Road Ahead. New Haven, CT: Yale University Press.

Romer, Paul M. 1990. Endogenous Technological Change. Journal of Political Economy 98(5, Part 2): S71–S102.

Romer, Paul. 1993."Idea Gaps and Object Gaps in Economic Development." Journal of Monetary Economics 32(3): 543–573.

Roosevelt, Franklin. 1936. Acceptance Speech for the Renomination for the Presidency, Philadelphia, June 27, 1936. In Samuel I. Rosenman (Ed.), *The Public Papers and Addresses of Franklin D. Roosevelt* (Vol. 5). New York: Random House, 230-236.

Roser, Max, and Hannah Ritchie. 2013. "Technological Progress." Our World in Data, 2013. https://ourworldindata.org/technological-progress

Rotemberg, Julio, and Michael Woodford. 1992. "Oligopolistic Pricing and the Effects of Aggregate Demand on Economic Activity." Journal of Political Economy 100(6): 1153-1207.

Rothstein, Mark, and Stacey Tovino. 2019. "California Takes the Lead on Data Privacy Law." Hastings Center Report 49(5): 4–5.

Rouffet. 1998. In re Rouffet. No. 97-1404. US Federal Circuit Court, July 15. https://law.justia.com/cases/federal/appellate-courts/F3/149/1350/560701/

Rowe, Frederick. 1962. Price Discrimination Under the Robinson-Patman Act. Boston: Little, Brown and Company.

RSNA. 2025. “Integrating Imaging Tools Helps Radiology AI Deliver Real Value.” RSNA News. https://www.rsna.org/news/2025/june/interoperability-helps-radiology-ai-deliver-value

Rubinfeld, Daniel. 2001. Antitrust Enforcement in Dynamic Network Industries. Antitrust Bulletin 46(3): 859-882.

Ruggie, John. 1982. International regimes, transactions, and change: Embedded liberalism in the postwar economic order. International Organization 36(2): 379–415.

Russell, Stuart, and Peter Norvig. 2021. Artificial Intelligence: A Modern Approach. 4th ed. Hoboken, NJ: Pearson.

Rysman, Marc. 2009. “The Economics of Two-Sided Markets.” Journal of Economic Perspectives 23(3): 125–143.

Rysman, Marc, and Julian Wright. 2014. The Economics of Payment Card Fee Structure: What Drives Payment Card Rewards? Review of Network Economics 13(3): 229-259.

S&P Dow Jones Indices 2025. S&P 500 Market Capitalization Concentration Report. https://www.spglobal.com/spdji/en/indices/equity/sp-500

Sabre 2020. United States v. Sabre Corp. et al., No. 1:19-cv-01548. U.S. District Court, D. Delaware, April 7. https://appliedantitrust.com/14_merger_litigation/cases_doj/sabre2019/2_d_del/sabre_ddel_opinion2020_04_08public_version.pdf

Sacra. 2025a. xAI revenue, valuation and funding. Sacra. https://sacra.com/c/xai/

Sacra. 2025b. Cohere revenue, valuation and funding. Sacra. https://sacra.com/c/cohere/

Saint-Martin, Léo. 2024. Economic Analysis of the EU and International Semiconductor Ecosystem. European Commission. https://ec.europa.eu/research/participants/documents/downloadPublic?appId=PPGMS&documentIds=080166e50f159e79

Sakagami, Ryo, Florian Lay, Andreas Dömel, Martin Schuster, Alin Albu-Schäffer, and Freek Stulp. 2023. Robotic world models—conceptualization, review, and engineering best practices. Frontiers in Robotics and AI 10:1253049.

Salemi, Alireza, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 7370–7392.

Salop, Steven. 2010. Question: What Is the Real and Proper Antitrust Welfare Standard—Answer: The Consumer Welfare Standard. Loyola University Chicago Law Journal 38:336.

Sampat, Bhaven. 2006. “Patenting and US Academic Research in the 20th Century.” Research Policy 35(6): 772–789.

Samuelson, Pamela, and Suzanne Scotchmer. 2002. "The Law and Economics of Reverse Engineering." Yale Law Journal 111(7): 1575–1663.

Sathish, Madhumitha, and Peter Rutten. 2024. Worldwide Artificial Intelligence Infrastructure Forecast, 2024–2028. IDC #US52597124. https://my.idc.com/getdoc.jsp?containerId=US52597124

Sato, Kaz, and Cliff Young. 2017. An in-depth look at Google’s first Tensor Processing Unit (TPU). Google Cloud Blog. https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

Savino, Vito. 2025. Edge computing and data centers in the age of AI. Data Center Dynamics. https://www.datacenterdynamics.com/en/opinions/edge-computing-and-data-centers-in-the-age-of-ai/

Scale AI. 2023. Scale Data Engine: Data Annotation, Collection, and Curation Platform. ScaleAI. https://scale.com/data-engine

ScaleAI. 2025. Scale AI Announces Next Phase of Company's Evolution. ScaleAI. https://scale.com/blog/scale-ai-announces-next-phase-of-company-evolution

Schaake, Marietje. 2021. "Regulating Technology with Marietje Schaake." WIRED. https://www.wired.com/story/regulating-technology

Schacht, Wendy H., and John Thomas. 2006. Patent Reform: Innovation Issues. CRS Report RL32996. https://www.everycrsreport.com/files/20061207_RL32996_3ee4011c8beb27223894b69fefc42a4c9b39377e.pdf

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? *Advances in Neural Information Processing Systems* 36: 27159–27182.

Schelling, Thomas. 1960. The Strategy of Conflict. Cambridge, MA: Harvard University Press.

Schmalensee, Richard. 2000. Antitrust Issues in Schumpeterian Industries. *American Economic Review* 90(2): 192–196.

Schmalensee, Richard. 2002. Payment Systems and Interchange Fees. *Journal of Industrial Economics* 50(2): 103–122.

Schmalensee, Richard. 1982. "Product Differentiation Advantages of Pioneering Brands." American Economic Review 72(3): 349-365.

Schmidt, Douglas. 2018. Google Data Collection. Nashville, TN: Vanderbilt University. https://www.dre.vanderbilt.edu/~schmidt/PDF/google-data-collection.pdf

Schneider, Bruno. 2026. Catch-up with the US or prosper below the tech frontier? An EU artificial intelligence strategy. Bruegel Policy Brief. https://www.bruegel.org/policy-brief/catch-us-or-prosper-below-tech-frontier-eu-artificial-intelligence-strategy

Schrage, Michael, and David Kiron. 2025. Philosophy eats AI. MIT Sloan Management Review. https://sloanreview.mit.edu/article/philosophy-eats-ai/

Schumpeter, Joseph. 1942. Capitalism, Socialism and Democracy. New York: Harper & Brothers.

Schwartz, Alan, and Robert Scott. 2002. "Contract Theory and the Limits of Contract Law." Yale Law Journal 113(3): 541-619.

Schwartz, Paul, and Edward Janger. 2002. The Gramm-Leach-Bliley Act, Information Privacy, and the Limits of Default Rules. *Minnesota Law Review* 86: 1217-1335.

Schwinn 1967. United States v. Arnold, Schwinn and Co., 388 U.S. 365. https://supreme.justia.com/cases/federal/us/388/365

Scoble, Robert. 2010. Walking through computer history with Apple cofounder Woz at the Computer History Museum. YouTube. https://www.youtube.com/watch?v=hsB8Hxnb52o

SCoJ. 1980. Patent and Trademark Law Amendments Act of 1980. Senate Report No. 96-480. 96th Congress, 2nd Session. Washington, DC: U.S. Government Printing Office.

Scribner, Marc. 2013. Slow Train Coming? Misguided Economic Regulation of U.S. Railroads, Then and Now. Issue Analysis No. 1. Competitive Enterprise Institute. https://cei.org/sites/default/files/Marc%20Scribner%20-%20Slow%20Train%20Coming.pdf

Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems* 28:2503-2511.

Seetharaman, Deepa, and Sriram Akash. 2025. "Nvidia to Invest Up to $100 Billion in OpenAI, Linking Two Artificial Intelligence Giants." Reuters. https://www.reuters.com/business/nvidia-invest-100-billion-openai-2025-09-22/

Seila Law. 2020. Seila Law LLC v. Consumer Financial Protection Bureau. 2020. No. 19-7. U.S. Supreme Court, June 29. https://supreme.justia.com/cases/federal/us/591/19-7/

Seliger, Yuri. 2025. Big Tech Financing Capacity and Balance Sheet Analysis. Bank of America Global Research, November 2025. https://investor.bankofamerica.com

Sell, Susan. 2003. Private Power, Public Law. Cambridge: Cambridge University Press.

SIA. 2020. 2020 State of the U.S. Semiconductor Industry Report. Washington, DC: Semiconductor Industry Association. https://www.semiconductors.org/wp-content/uploads/2020/07/2020-SIA-State-of-the-Industry-Report-FINAL-1.pdf

SIA. 2021. 2021 State of the U.S. Semiconductor Industry Report. Washington, DC: Semiconductor Industry Association. https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf

Serant, Claire. 2000. "Why Motorola Chose Outsourcing Route." EE Times. https://www.eetimes.com/why-motorola-chose-outsourcing-route/

Setser, Brad. 2020. "When the Services Trade Data Tells You More About Tax Avoidance Than About Actual Trade." Council on Foreign Relations Blog. https://www.cfr.org/blog/when-services-trade-data-tells-you-more-about-tax-avoidance-about-actual-trade

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn and Pablo Villalobos. 2022. Compute Trends Across Three Eras of Machine Learning. arXiv preprint arXiv:2202.05924. https://arxiv.org/abs/2202.05924

Shakudo. 2025. “12 MLOps Best Practices Every Enterprise Needs in 2025.” Shakudo Blog. https://www.shakudo.io/blog/mlops-best-practices-enterprise-2025

Shamsuddoha, Mohammad, Tasnuba Nasir, and Mohammad Saifuddoha Fawaaz. 2025. “Humanoid Robots like Tesla Optimus and the Future of Supply Chains: Enhancing Efficiency, Sustainability, and Workforce Dynamics.” Automation 6(1): 9.

Shankland, Stephen. 2004. "Intel Shifts 64-bit Emphasis." CNET. https://www.cnet.com/tech/tech-industry/intel-shifts-64-bit-emphasis

Shapira, Philip. 2001. “US Manufacturing Extension Partnerships.” Research Policy 30(6): 977–992.

Shapiro, Carl, and Hal Varian. 1999. Information Rules: A Strategic Guide to the Network Economy. Boston, MA: Harvard Business School Press.

Shapiro, Carl. 1989. "Theories of Oligopoly Behavior." In Handbook of Industrial Organization, Vol. 1, edited by Richard Schmalensee and Robert Willig. Amsterdam: Elsevier, 329-414.

Shapiro, Carl. 1996. "Mergers with Differentiated Products." Antitrust 10(2): 23–30.

Shapiro, Carl. 2010a. "The Role of Antitrust in Preventing Patent Holdup." Antitrust Law Journal 82(2): 201–275.

Shapiro, Carl. 2010b. “The 2010 Horizontal Merger Guidelines: From Hedgehog to Fox in Forty Years.” Antitrust Law Journal 77: 49–168.

Shapiro, Carl. 2018. “Competition and the Dynamics of Industrial Structure.” In Innovation Policy and the Economy, Vol. 18, edited by Josh Lerner and Scott Stern. Chicago: University of Chicago Press, 23–63.

Shapiro, Steven. 1993. “Preliminary Injunction Motions in Patent Litigation.” IDEA: The Journal of Law and Technology 33: 323–348.

Sharma, Debendra Das and Ravi Mahajan. 2024. Advanced packaging of chiplets for future computing needs. Nature Electronics 7(3): 181–192.

Shelanski, Howard. 2011. Enforcing Competition During an Economic Emergency. Minnesota Law Review 95:1459.

Shelanski, Howard, and Gregory Sidak. 2001. "Antitrust Divestiture in Network Industries." University of Chicago Law Review 68(1): 1–99.

Shepherd, George. 1996. "Fierce Compromise." Northwestern University Law Review 90(4): 1557–1683.

Shojaee, Parshin, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. “The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity.” airXiv preprint airXiv: 2506.06941. https://arxiv.org/pdf/2506.06941

Shumailov, Ilia, Yarin Gal, Vitaly Kurin, Nicholas Lane, and Ross Anderson. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv preprint: arXiv.2305.17493. https://doi.org/10.48550/arXiv.2305.17493

Sidak, Gregory and David Teece. 2009. "Dynamic Competition in Antitrust Law." Journal of Competition Law and Economics 5(4): 581–631.

Sidak, Gregory. 2013. "The Meaning of FRAND, Part I: Royalties." Journal of Competition Law and Economics 9(4): 931–1055.

Singhal, Karan, et al. 2023. Large language models encode clinical knowledge. Nature 620: 172-180.

Sites, Richard, and Richard Witek. 1993. Alpha AXP Architecture Reference Manual. 2nd ed. Burlington, MA: Digital Press.

Skibba, Ramin, Subhasish Mitra, and Philip Wong. 2015. “Skyscraper-Style Chip Design Boosts Performance by Factor of a Thousand.” Stanford Report. https://biox.stanford.edu/highlight/skyscraper-style-chip-design-boosts-performance-factor-thousand

Skidmore et al. 1944. Skidmore et al. v. Swift and Co., No. 12, U.S. Supreme Court, December 4. https://tile.loc.gov/storage-services/service/ll/usrep/usrep323/usrep323134/usrep323134.pdf

Slack Developers. 2025. Rate limit changes for non-Marketplace apps. Slack Developers. https://docs.slack.dev/changelog/2025/05/29/rate-limit-changes-for-non-marketplace-apps/

Smith, Brad 2025. The Golden Opportunity for American AI. Microsoft Blog. https://blogs.microsoft.com/on-the-issues/2025/01/03/the-golden-opportunity-for-american-ai/

Smith, Ben. 2018. "Project Strobe: Protecting Your Data, Improving Our Third-Party APIs, and Sunsetting Consumer Google+." Google Blog. https://blog.google/technology/safety-security/project-strobe/

Smith, Bob. 2024. Semiconductor Engineering. “The State of the EDA Industry in 2024.” https://semiengineering.com/the-state-of-the-eda-industry-in-2024/

Smith, Steven, Mike Hanlon, and Robert Bailey. 1992. Power Management for a Laptop Computer with Slow and Sleep Modes. (US Patent No. 5,167,024). USPTO. https://patents.google.com/patent/US5167024A/en

Snowflake.help. 2025. How Snowflake's Open Semantic Interchange is revolutionizing AI data readiness. Snowflake.help. https://snowflake.help/how-snowflakes-open-semantic-interchange-is-revolutionizing-ai-data-readiness/

Socony-Vacuum Oil. 1940. United States v. Socony-Vacuum Oil Co., 310 U.S. 150. https://supreme.justia.com/cases/federal/us/310/150

SoftBank Group Corp. 2025. SoftBank Group Report 2025 (Annual Report). https://group.softbank/en/news/info/20250728

Sokol, Daniel, and Roisin Comerford. 2020. Does Antitrust Have a Role to Play in Regulating Big Data? In *The Cambridge Handbook of Antitrust, Intellectual Property, and High Tech*, edited by Roger Blair and D. Daniel Sokol. Cambridge: Cambridge University Press, 271–291.

Solove, Daniel, and Woodrow Hartzog. 2014. "The FTC and the New Common Law of Privacy." Columbia Law Review 114(3): 583–676.

Solow, Robert. 1988. Growth Theory and After. *The American Economic Review* 78(3): 307–317.

Solow, Robert. 1957. "Technical Change and the Aggregate Production Function." Review of Economics and Statistics 39(3): 312-320.

Solow, Robert. 1987. "We'd Better Watch Out." The New York Times Book Review, July 12, 1987, p. 36.

Solsman, Joan. 2018. "YouTube's AI Is the Puppet Master over Most of What You Watch." CNET. https://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan/

Son, Hugh. 2025. "Goldman Sachs Launches AI Assistant as the Tech Sweeps Banking." CNBC. https://www.cnbc.com/2025/01/21/goldman-sachs-launches-ai-assistant.html

Song, Seagull and Mo Wang. 2024. “China’s First Generative AI Copyright Infringement Case: Ultraman.” King and Wood Mallesons. https://www.kwm.com/cn/en/insights/latest-thinking/china-s-first-case-on-aigc-output-infringement-ultraman.html

Soper, Taylor 2024. Databricks, set to raise massive $10B funding round, looks to grow Seattle-area footprint. GeekWire. https://www.geekwire.com/2024/databricks-set-to-raise-massive-10b-funding-round-looks-to-grow-seattle-area-footprint

Specht et al. 2022. Christopher Specht et al. v. Netscape Communications Corporation and America Online, Inc., Nos. 01-7870, 01-7860, U.S. Court of Appeals, Second Circuit, October 1. https://scholar.google.com/scholar_case?case=9587085159184835436

Spectrum Sports. 1993. Spectrum Sports, Inc. v. McQuillan, 506 U.S. 447, U.S. Supreme Court, January 13. https://supreme.justia.com/cases/federal/us/506/447

Spegele, Brian. 2025. “The Fortress That China Built for Its Battle With America,” Wall Street Journal. https://www.wsj.com/world/china/china-us-technology-economy-advancements-bb8d7439

Spence, Michael. 2021. "Some Thoughts on the Washington Consensus and Subsequent Global Development Experience." Journal of Economic Perspectives 35(3): 67–82.

Spengler, Joseph. 1950. "Vertical Integration and Antitrust Policy." Journal of Political Economy 58(4): 347–352.

Spulber, Daniel. 2015. "How Patents Provide the Foundation of the Market for Inventions." Journal of Competition Law and Economics 11(2): 271-316.

Srinivasan, Dina. 2019. “The Antitrust Case Against Facebook.” Berkeley Business Law Journal 16(1): 39–98.

Srivathsan, Bhargs, Marc Sorel, and Pankaj Sachdeva, with Arjita Bhan, Haripreet Batra, Raman Sharma, Rishi Gupta, and Surbhi Choudhary. 2024. "Power to the People: The Gigawatt-Scale Challenge of AI Data Centers." McKinsey & Company Insights. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand#/

Stainless. 2025. Production-ready MCP servers optimized for agentic coding and context limits. https://www.stainless.com/products/mcp

Stalk, George, Philip Evans, and Lawrence Shulman. 1992. "Competing on Capabilities: The New Rules of Corporate Strategy." Harvard Business Review 70(2): 57-69.

Standard Oil. 1911. Standard Oil Co. of New Jersey v. United States, 221 U.S. 1, U.S. Supreme Court, May 15. https://supreme.justia.com/cases/federal/us/221/1/

Standard Oil. 1949. Standard Oil Co. of California v. United States, No. 279. US Supreme Court, June 13. https://supreme.justia.com/cases/federal/us/337/293

Staples. 1997. FTC v. Staples, Inc., 970 F. Supp. 1066 (D.D.C.). https://law.justia.com/cases/federal/district-courts/FSupp/970/1066/1639260

Stasik, Eric. 2010. "Royalty Rates and Licensing Strategies for Essential Patents on LTE (4G) Telecommunication Standards." Les Nouvelles 45(3): 114–119.

StatCounter. 2024. Mobile operating system market share worldwide. StatCounter Global Stats. https://gs.statcounter.com/os-market-share/mobile/worldwide

StatCounter. 2025. “Browser Market Share Worldwide.” StatCounter Global Stats. https://gs.statcounter.com/browser-market-share/

State Street Bank. 1998. State Street Bank and Trust Co. v. Signature Financial Group, Inc. No. 96-1327. US Federal Circuit Court, July 23. https://law.justia.com/cases/federal/appellate-courts/F3/149/1368/560460

State Street Bank. 1998. State Street Bank and Trust Co. v. Signature Financial Group, Inc. No. 96-1327. US Federal Circuit Court, July 23. https://law.justia.com/cases/federal/appellate-courts/F3/149/1368/560460/

Statista. 2024a. "Volume of Data/Information Created, Captured, Copied, and Consumed Worldwide from 2010 to 2020, with Forecasts from 2021 to 2025." Statista. https://www.statista.com/statistics/871513/worldwide-data-created

Statista. 2024b. Leading ad-supported digital video platforms in the United States, by share of viewing time. Statista. https://www.statista.com/statistics/276623/share-of-video-streaming-usage-in-the-us-by-service/

Stein, David, Kaitlin Johns, Maryam Fatima, and Shari Allen. 2024. U.S. International Services: Trade in Services in 2023 and Services Supplied Through Affiliates in 2022. U.S. Bureau of Economic Analysis. https://apps.bea.gov/scb/issues/2024/10-october/1024-international-services.htm

Stern, Jonathan. 2001. The Electronic Signatures in Global and National Commerce Act. Berkeley Technology Law Journal 16(1): 391-414.

Stevens, Ashley. 2004. "The Enactment of Bayh-Dole." Journal of Technology Transfer 29(1): 93–99.

Stigler, George. 1968. The Organization of Industry. Homewood, IL: Richard D. Irwin.

Stigler, George. 1971. "The Theory of Economic Regulation." Bell Journal of Economics and Management Science 2(1): 3-21.

Stiglitz, Joseph. 2002. Globalization and Its Discontents. New York: W.W. Norton and Company.

Stiglitz, Joseph. 2006. Making Globalization Work. New York: W.W. Norton and Company.

Stocking, George, and Willard Mueller. 1955. The Cellophane Case and the New Competition. The American Economic Review 45(1): 29-63.

Stranger, Greg, and Shane Greenstein. 2007. "Pricing at the On-Ramp to the Internet: Price Indexes for ISPs during the 1990s." In Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches, edited by Ernst Berndt and Charles Hulten, Chicago: University of Chicago Press, 197–234.

Strauss, Robert. 1979. “Symposium–The Tokyo Round: It’s Meaning and Effect, Introduction.” Georgia Journal of International and Comparative Law 9(2): 151–175.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

Studiengesellschaft Kohle 1981. U.S. v. Studiengesellschaft Kohle et al., No. 79-1634, U.S. Court of Appeals, District of Columbia Circuit, December 11. https://scholar.google.com/scholar_case?case=867784977804525218

Stuecke, Jan. 2025. Top 7 Open-Source Frameworks for Federated Learning. Apheris. https://www.apheris.com/resources/blog/top-7-open-source-frameworks-for-federated-learning

Sukumar, Vinesh. 2025. “Shifting AI inference from the cloud to your phone can reduce AI costs.” Qualcomm OnQ Blog. https://www.qualcomm.com/news/onq/2025/09/shifting-ai-inference-from-the-cloud-to-your-phone-can-reduce-ai-costs

Suleyman, Mustafa. 2023. The Coming Wave: Technology, Power, and the Twenty-first Century's Greatest Dilemma. New York: Crown. https://the-coming-wave.com

Sullivan and Cromwell LLP. 2020. "Ninth Circuit Holds That Qualcomm's Patent Licensing Program Does Not Violate U.S. Antitrust Law." Sullivan and Cromwell. https://www.sullcrom.com/SullivanCromwell/_Assets/PDFs/Memos/sc-publication-ninth-circuit-holds-qualcomm-patent-licensing-program-does-not-violate-us-antitrust-law.pdf

Sullivan, Lawrence, and Warren Grimes. 2006. The Law of Antitrust. 2nd ed. St. Paul: West Academic Publishing.

Sullivan, Nolan. 2025. In Depth: Speakeasy vs Stainless. Speakeasy Blog. https://www.speakeasy.com/blog/speakeasy-vs-stainless

Sunstein, Cass. 2002. The Cost-Benefit State: The Future of Regulatory Protection. Chicago, IL: American Bar Association.

Sunstein, Cass. 2013. "The Office of Information and Regulatory Affairs." Harvard Law Review 126(7): 1838–1878.

Surapaneni, Rao, Miku Jha, Michael Vakoc, and Todd Segal. 2025. Announcing the Agent-to-Agent (A2A) protocol: A new era of agent interoperability. Google Developers Blog. https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/

Svensson, Philip. 2024. Qualcomm: Quality communication and semiconductor innovation. Quartr. https://quartr.com/insights/company-research/qualcomm-quality-communication-and-semiconductor-innovation

SWIFT. 2025. Momentum builds as industry advances ISO 20022 adoption. https://www.swift.com/news-events/news/momentum-builds-industry-advances-iso-20022-adoption

Synergy Research Group. 2024. "Cloud Infrastructure Market Share: AWS, Microsoft and Google Control 67%." Synergy Research Group. https://www.prnewswire.com/news-

releases/cloud-market-gets-its-mojo-back-ai-helps-push-q4-increase-in-cloud-spending-to-new-highs-302051366.html

Syufy Enterprises. 1990. United States v. Syufy Enterprises, 903 F.2d 659, U.S. Court of Appeals for the Ninth Circuit, May 9. https://law.justia.com/cases/federal/appellate-courts/F2/903/659/435777

Syverson, Chad. 2011. What Determines Productivity? *Journal of Economic Literature* 49(2): 326–365.

Sze, Simon, and Kwok Ng. 2006. Physics of Semiconductor Devices. 3rd ed. Hoboken, NJ: John Wiley and Sons.

Tan, Lisa, and Ken Huang. 2025. “The AI Agent Economy.” In Progress in Information Systems, edited by Jörg Becker and Ralf Knackstedt. Cham: Springer Nature: 99–134.

Tao, Fei, Qinglin Qi, Ang Liu, and Andrew Kusiak. 2018. Data-driven smart manufacturing. Journal of Manufacturing Systems 48(Part C): 157–169.

Tassey, Gregory. 2000. "Standardization in Technology-Based Markets." Research Policy 29(4-5): 587–602.

Teare, Gené. 2025. Startup Funding Regained Its Footing In 2024 As AI Became the Star of the Show. Crunchbase News, January 7. https://news.crunchbase.com/venture/global-funding-data-analysis-ai-eoy-2024/

Techlevated. 2023. FinFET vs MOSFET (Planar) Transistors in Chip Design. Techlevated. https://techlevated.com/finfet-vs-mosfet-planar-transistor/

Teece, David. 1982. "Towards an Economic Theory of the Multiproduct Firm." Journal of Economic Behavior and Organization 3(1): 39-63.

Teece, David. 1984. "Economic Analysis and Strategic Management." California Management Review 26 (3): 87–110.

Tellex, Stefanie, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. Annual Review of Control, Robotics, and Autonomous Systems 3: 25–55.

Temin, Peter and Louis Galambos. 1987. The Fall of the Bell System. New York: Cambridge University Press.

Temkin, Marina. 2025. Perplexity reportedly raised $200M at $20B valuation. TechCrunch. https://techcrunch.com/2025/09/10/perplexity-reportedly-raised-200m-at-20b-valuation/

Tempus. 2025. "Clinical trial matching (Tempus TIME)." Tempus. https://www.tempus.com/oncology/clinical-trial-matching/

Tene, Omer, Bethany Withers, and Tayjus Surampudi. 2025. California Moves to Regulate Frontier AI With a Focus on Catastrophic Risk. *Goodwin Law Alert*. https://www.goodwinlaw.com/en/insights/publications/2025/11/alerts-technology-aiml-california-moves-to-regulate-frontier-ai-with-a-focus-on-catastrophic-risk

Teti, Feodora. 2020. 30 Years of Trade Policy: Evidence from 5.7 Billion Tariffs. ifo Working Papers No. 334. https://www.ifo.de/DocDL/wp-2020-334-teti-trade-policy-tariffs.pdf

Texas Instruments. 2021. Developing the First ICs to Orbit Earth. TI Blog. https://www.ti.com/about-ti/newsroom/company-blog/developing-the-first-ics-to-orbit-earth.html

The Economist. 2002 Innovation’s golden goose. The Economist. December 14: 3.

The Economist. 2015a. A tightening grip. https://www.economist.com/briefing/2015/03/12/a-tightening-grip

The Economist. 2015b. Made in China? https://www.economist.com/leaders/2015/03/12/made-in-china

The Economist. 2017. "The World's Most Valuable Resource Is No Longer Oil, but Data." The Economist. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

The Economist. 2020. Are data more like oil or sunlight? https://www.economist.com/special-report/2020/02/20/are-data-more-like-oil-or-sunlight

The Information Staff. 2024. 7 Charts That Explain 2024. The Information. https://www.theinformation.com/articles/7-charts-that-explain-2024

Thompson, Alan. 2024. Models: Gemini. LifeArchitect.ai. https://lifearchitect.ai/gemini/

Tierno, Paul. 2024. Artificial intelligence and machine learning in financial services. CRS Report R47997. https://www.congress.gov/crs-product/R47997

Tilley, Aaron. 2024. “How Apple Fell Behind in the AI Arms Race.” The Wall Street Journal. https://www.wsj.com/tech/ai/apple-ai-siri-development-behind-9ea65ee8

Tirole, Jean. 1988. The Theory of Industrial Organization. Cambridge, MA: MIT Press.

Topco Associates. 1972. United States v. Topco Associates, Inc., 405 U.S. 596. https://supreme.justia.com/cases/federal/us/405/596

Torchia, Marcus and Shirer, Matt 2024. New IDC Spending Guide Forecasts Edge Computing Investments to Reach $232 Billion in 2024. IDC. https://www.businesswire.com/news/home/20240314632794/en/New-IDC-Spending-Guide-Forecasts-Edge-Computing-Investments-Will-Reach-%24232-Billion-in-2024

Touvron, Hugo, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971. https://arxiv.org/abs/2302.13971

Train, Kenneth. 2009. Discrete Choice Methods with Simulation. 2nd ed. Cambridge: Cambridge University Press.

Trenton Potteries. 1927. United States v. Trenton Potteries Co., 273 U.S. 392, U.S. Supreme Court, February 21. https://supreme.justia.com/cases/federal/us/273/392

Trump, Donald. 2017. " Executive Order 13771—Reducing Regulation and Controlling Regulatory Costs. 82 FR: 9339. https://www.govinfo.gov/app/details/FR-2017-02-03/2017-02451

Trump, Donald. 2025a. America's AI Action Plan—Winning the Race. White House. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

Trump, Donald. 2025b. Executive Order 14365. "Ensuring a National Policy Framework for Artificial Intelligence." Federal Register 90, no. 240: 58499. https://www.federalregister.gov/documents/2025/12/16/2025-23092/ensuring-a-national-policy-framework-for-artificial-intelligence

Trump, Donald. 2025c. "Executive Order 14257—Regulating Imports With a Reciprocal Tariff to Rectify Trade Practices That Contribute to Large and Persistent Annual United States Goods Trade Deficits." https://www.whitehouse.gov/presidential-actions/2025/04/regulating-imports-with-a-reciprocal-tariff-to-rectify-trade-practices-that-contribute-to-large-and-persistent-annual-united-states-goods-trade-deficits/

Tschand, Arya, et al. 2024. "MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from Microwatts to Megawatts for Sustainable AI." arXiv preprint arXiv: 2410.12032. https://arxiv.org/abs/2410.12032

TSMC. 2023a. Annual Report. Hsinchu, Taiwan: TSMC (Investor Relations). https://www.tsmc.com/english/investorRelations/annual_reports

TSMC. 2023b. Sustainability Report. Hsinchu, Taiwan: TSMC (ESG Reporting). https://esg.tsmc.com/en/

TSMC. 2024. TSMC Annual Report 2024. TSMC. https://investor.tsmc.com/sites/ir/annual-report/2024/2024%20Annual%20Report.E.pdf

Tsui, David, and Andrew Chang. 2025. Research Update: NVIDIA Corp. Outlook Revised to Positive. S&P Global Ratings. https://www.spglobal.com/ratings/en/regulatory/article/-/view/type/HTML/id/3463382

Turner, Donald. 1949. The Definition of Agreement Under the Sherman Act: Conscious Parallelism and Refusals to Deal. Harvard Law Review 62(4): 661-701.

Tyman, Annette 2024. Mobley v. Workday: Court holds AI service providers could be directly liable for employment discrimination. Seyfarth Shaw LLP. https://www.seyfarth.com/news-insights/mobley-v-workday-court-holds-ai-service-providers-could-be-directly-liable-for-employment-discrimination-under-agent-theory.html

US Census Bureau. n.d. Exports from Manufacturing Establishments: Historical Data. DoC. https://www.census.gov/manufacturing/exports/historical_data/index.html

US Congress. 1890. Sherman Act. Ch. 647, 26 Stat. 209. https://www.govinfo.gov/content/pkg/COMPS-3055/pdf/COMPS-3055.pdf

US Congress. 1914a. Clayton Act. Ch. 323, 38 Stat. 730. https://www.govinfo.gov/content/pkg/COMPS-3049/pdf/COMPS-3049.pdf

US Congress. 1914b. Federal Trade Commission Act. Ch. 311, 38 Stat. 717. https://www.ftc.gov/legal-library/browse/statutes/federal-trade-commission-act

US Congress. 1934. Communications Act of 1934, Pub. L. No. 73–416. https://www.govinfo.gov/content/pkg/STATUTE-48/pdf/STATUTE-48-Pg1064.pdf

US Congress. 1936. Robinson-Patman Antidiscrimination Act. Ch. 592, 49 Stat. 1526. https://www.govinfo.gov/content/pkg/COMPS-12152/pdf/COMPS-12152.pdf

US Congress. 1946. Administrative Procedure Act. Pub. L. No. 79-404. https://www.govinfo.gov/content/pkg/STATUTE-60/pdf/STATUTE-60-Pg237.pdf

US Congress. 1950. Celler-Kefauver Act. Pub. L. No. 81-899, 64 Stat. 1125. https://www.govinfo.gov/content/pkg/STATUTE-64/pdf/STATUTE-64-Pg1125.pdf

US Congress. 1952. Patent Act of 1952. Pub. L. No. 82-593. https://www.govinfo.gov/content/pkg/STATUTE-66/pdf/STATUTE-66-Pg792.pdf

US Congress. 1965. Food and Agriculture Act of 1965. Pub. L. No. 89-321. https://nationalaglawcenter.org/wp-content/uploads/assets/farmbills/1965.pdf

US Congress. 1976a. Hart-Scott-Rodino Antitrust Improvements Act. Pub. L. No. 94-435, 90 Stat. 1383. https://www.govinfo.gov/content/pkg/STATUTE-90/pdf/STATUTE-90-Pg1383.pdf

US Congress. 1976b. US Copyright Act. Pub. L. 94–553. https://www.copyright.gov/history/pl94-553.pdf

US Congress. 1978. Public Utility Regulatory Policies Act of 1978. Pub. L. No. 95-617. https://www.congress.gov/bill/95th-congress/house-bill/4018

US Congress. 1979a. The 1979 Joint Economic Report. 96th Congress, 1st Session. https://www.jec.senate.gov/reports/96th%20Congress/The%201979%20Joint%20Economic%20Report%20(930).pdf

US Congress. 1979b. Trade Agreements Act of 1979. Pub. L. No. 96-39. https://www.congress.gov/bill/96th-congress/house-bill/4537

US Congress. 1980a. Stevenson-Wydler Technology Innovation Act of 1980, Pub. L. No. 96-480. https://uscode.house.gov/statutes/pl/96/480.pdf

US Congress. 1980b. Patent and Trademark Law Amendments Act (Bayh-Dole Act), Pub. L. No. 96-517. https://www.govinfo.gov/content/pkg/STATUTE-94/pdf/STATUTE-94-Pg3015.pdf

US Congress. 1981. H.R. Rep. No. 97-312, 97th Cong., 1st Sess. (Federal Courts Improvement Act of 1981). https://www.congress.gov/congressional-report/97th-congress/house-report/312

US Congress. 1982. Federal Courts Improvement Act of 1982. Pub. L. No. 97-164. https://www.congress.gov/bill/97th-congress/house-bill/4482

US Congress. 1984a. Semiconductor Chip Protection Act of 1984, Pub. L. No. 98-620. https://www.congress.gov/98/statute/STATUTE-98/STATUTE-98-Pg3335.pdf

US Congress. 1984b. National Cooperative Research Act of 1984. Pub. L. No. 98-462. https://www.congress.gov/bill/98th-congress/senate-bill/1841

U.S. Congress. 1986a. Federal Technology Transfer Act of 1986. Pub. L. No. 99–502. https://www.congress.gov/bill/99th-congress/house-bill/3773

US Congress. 1986b. Computer Fraud and Abuse Act of 1986. Pub. L. No. 99-474. https://www.congress.gov/bill/99th-congress/house-bill/4718

US Congress. 1988. Omnibus Trade and Competitiveness Act of 1988. Pub. L. No. 100-418. https://www.congress.gov/bill/100th-congress/house-bill/4848

US Congress. 1996. Telecommunications Act of 1996. Pub. L. 104–104. https://www.congress.gov/104/plaws/publ104/PLAW-104publ104.pdf

US Congress. 1998. Digital Millennium Copyright Act. Pub. L. 105–304. https://www.congress.gov/105/plaws/publ304/PLAW-105publ304.pdf

US Congress. 1999. Gramm–Leach–Bliley Act. Pub. L. No. 106. https://www.govinfo.gov/content/pkg/PLAW-106publ102/pdf/PLAW-106publ102.pdf.

US Congress. 2000. Electronic Signatures in Global and National Commerce Act (E-SIGN Act). Pub. L. 106-229. https://www.congress.gov/bill/106th-congress/house-bill/1714

US Congress. 2002. E-Government Act of 2002. Pub. L. No. 107-347. https://www.congress.gov/bill/107th-congress/house-bill/2458

US Congress. 2004. Standards Development Organization Advancement Act of 2004. Pub. L. No. 108-237. https://www.congress.gov/bill/108th-congress/house-bill/1086

US Congress. 2016. Defend Trade Secrets Act of 2016. Pub. L. No. 114-153. https://www.congress.gov/114/plaws/publ153/PLAW-114publ153.pdf

US Congress. 2018. Allow States and Victims to Fight Online Sex Trafficking Act of 2017. Pub. L. No. 115-164. https://www.congress.gov/bill/115th-congress/house-bill/1865

US Congress. 2022. CHIPS and Science Act of 2022. Pub. L. 117–167. https://uscode.house.gov/download/bills/117-2/117-167.pdf

USPTO. 2013. Policy Statement on Remedies for Standards-Essential Patents Subject to Voluntary F/RAND Commitments. Washington, DC.

USPTO. 2021. U.S. Patent Statistics Chart Calendar Years 1963-2020. https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm

USPTO. 2024. MPEP. 9th ed. DoC. https://www.uspto.gov/web/offices/pac/mpep/index.html

Ugwi, Monica. 2025. “The Future of Manufacturing: AI for Data Standardization.” Microsoft Industry Blogs. https://www.microsoft.com/en-us/industry/blog/manufacturing-and-mobility/manufacturing/2025/01/29/the-future-of-manufacturing-ai-for-data-standardization/

UK CaMA. 2020. Online Platforms and Digital Advertising. UK CaMA. https://assets.publishing.service.gov.uk/media/5fa557668fa8f5788db46efc/Final_report_Digital_ALT_TEXT.pdf

CMA. 2020. Online Platforms and Digital Advertising Market Study. London: Competition and Markets Authority.

ULC. 1999. Uniform Electronic Transactions Act (UETA). National Conference of Commissioners on Uniform State Laws. https://www.uniformlaws.org/viewdocument/final-act-21?CommunityKey=2c04b76c-2b7d-4399-977e-d5876ba7e034.

UN (United Nations). 2019. Globally Harmonized System of Classification and Labelling of Chemicals (GHS). UN. https://unece.org/ghs-rev8-2019

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. 2025. Guidance for Generative AI in Education and Research (Updated from the 2023 version). Paris: UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Unify. 2025a. Employee data and trends for OpenAI. https://www.unifygtm.com/insights-headcount/openai

Unify. 2025b. Employee data and trends for Cohere. https://www.unifygtm.com/insights-headcount/cohere

Urbain, Thomas, and Luna Lin. 2025. As US battles China on AI, some companies choose Chinese. TechXplore. https://techxplore.com/news/2025-12-china-ai-companies-chinese.html

Urofsky, Melvin I. 2009. Louis D. Brandeis: A Life. New York: Pantheon Books.

US Congress. 1890. Sherman Act. Ch. 647, 26 Stat. 209. https://www.govinfo.gov/content/pkg/COMPS-3055/pdf/COMPS-3055.pdf

US Congress. 1914a. Clayton Act. Ch. 323, 38 Stat. 730. https://www.govinfo.gov/content/pkg/COMPS-3049/pdf/COMPS-3049.pdf

US Congress. 1914b. Federal Trade Commission Act. Ch. 311, 38 Stat. 717. https://www.ftc.gov/legal-library/browse/statutes/federal-trade-commission-act

US Congress. 1934. Communications Act of 1934, Pub. L. No. 73–416.
https://www.govinfo.gov/content/pkg/STATUTE-48/pdf/STATUTE-48-Pg1064.pdf
US Congress. 1936. Robinson-Patman Antidiscrimination Act. Ch. 592, 49 Stat. 1526.
https://www.govinfo.gov/content/pkg/COMPS-12152/pdf/COMPS-12152.pdf
US Congress. 1946. Administrative Procedure Act. Pub. L. No. 79-404.
https://www.govinfo.gov/content/pkg/STATUTE-60/pdf/STATUTE-60-Pg237.pdf
US Congress. 1950. Celler-Kefauver Act. Pub. L. No. 81-899, 64 Stat. 1125.
https://www.govinfo.gov/content/pkg/STATUTE-64/pdf/STATUTE-64-Pg1125.pdf
US Congress. 1952. Patent Act of 1952. Pub. L. No. 82-593.
https://www.govinfo.gov/content/pkg/STATUTE-66/pdf/STATUTE-66-Pg792.pdf
US Congress. 1965b. Food and Agriculture Act of 1965. Pub. L. No. 89-321.
https://nationalaglawcenter.org/wp-content/uploads/assets/farmbills/1965.pdf
US Congress. 1976a. Hart-Scott-Rodino Antitrust Improvements Act. Pub. L. No. 94-435, 90 Stat. 1383. https://www.govinfo.gov/content/pkg/STATUTE-90/pdf/STATUTE-90-Pg1383.pdf
US Congress. 1976b. US Copyright Act. Pub. L. 94–553.
https://www.copyright.gov/history/pl94-553.pdf
US Congress. 1978. Public Utility Regulatory Policies Act of 1978. Pub. L. No. 95-617.
https://www.congress.gov/bill/95th-congress/house-bill/4018
US Congress. 1979a. The 1979 Joint Economic Report. 96th Congress, 1st Session.
https://www.jec.senate.gov/reports/96th%20Congress/The%201979%20Joint%20Economic%20Report%20(930).pdf
US Congress. 1979b. Trade Agreements Act of 1979. Pub. L. No. 96-39.
https://www.congress.gov/bill/96th-congress/house-bill/4537

US Congress. 1980a. Patent and Trademark Law Amendments Act (Bayh-Dole Act), Pub. L. No. 96-517. https://www.govinfo.gov/content/pkg/STATUTE-94/pdf/STATUTE-94-Pg3015.pdf

US Congress. 1980b. Stevenson-Wydler Technology Innovation Act of 1980, Pub. L. No. 96-480. https://uscode.house.gov/statutes/pl/96/480.pdf

US Congress. 1981. H.R. Rep. No. 97-312, 97th Cong., 1st Sess. (Federal Courts Improvement Act of 1981). https://www.congress.gov/congressional-report/97th-congress/house-report/312

US Congress. 1982. Federal Courts Improvement Act of 1982. Pub. L. No. 97-164. https://www.congress.gov/bill/97th-congress/house-bill/4482

US Congress. 1984a. Semiconductor Chip Protection Act of 1984, Pub. L. No. 98-620. https://www.congress.gov/98/statute/STATUTE-98/STATUTE-98-Pg3335.pdf

US Congress. 1984b. National Cooperative Research Act of 1984. Pub. L. No. 98-462. https://www.congress.gov/bill/98th-congress/senate-bill/1841

US Congress. 1986b. Computer Fraud and Abuse Act of 1986. Pub. L. No. 99-474. https://www.congress.gov/bill/99th-congress/house-bill/4718

US Congress. 1988. Omnibus Trade and Competitiveness Act of 1988. Pub. L. No. 100-418. https://www.congress.gov/bill/100th-congress/house-bill/4848

US Congress. 1996. Telecommunications Act of 1996. Pub. L. 104–104. https://www.congress.gov/104/plaws/publ104/PLAW-104publ104.pdf

US Congress. 1998. Digital Millennium Copyright Act. Pub. L. 105–304. https://www.congress.gov/105/plaws/publ304/PLAW-105publ304.pdf

US Congress. 1999. Gramm–Leach–Bliley Act. Pub. L. No. 106. https://www.govinfo.gov/content/pkg/PLAW-106publ102/pdf/PLAW-106publ102.pdf.

US Congress. 2000. Electronic Signatures in Global and National Commerce Act (E-SIGN Act). Pub. L. 106-229. https://www.congress.gov/bill/106th-congress/house-bill/1714

US Congress. 2002. E-Government Act of 2002. Pub. L. No. 107-347. https://www.congress.gov/bill/107th-congress/house-bill/2458

US Congress. 2004. Standards Development Organization Advancement Act of 2004. Pub. L. No. 108-237. https://www.congress.gov/bill/108th-congress/house-bill/1086

US Congress. 2016. Defend Trade Secrets Act of 2016. Pub. L. No. 114-153. https://www.congress.gov/114/plaws/publ153/PLAW-114publ153.pdf

US Congress. 2018. Allow States and Victims to Fight Online Sex Trafficking Act of 2017. Pub. L. No. 115-164. https://www.congress.gov/bill/115th-congress/house-bill/1865

US Congress. 2022. CHIPS and Science Act of 2022. Pub. L. 117–167. https://uscode.house.gov/download/bills/117-2/117-167.pdf

USPTO. 1995. "Request for Comments on Proposed Examination Guidelines for Computer-Implemented Inventions." Federal Register 60 (106): 28778–28780.

USPTO. 1996. "Examination Guidelines for Computer-Related Inventions." Federal Register 61(40): 7478–7492.

USPTO. 2020. "Public Views on Artificial Intelligence and Intellectual Property Policy." U.S. Patent and Trademark Office, October 6. https://www.uspto.gov/subscription-center/2020/uspto-releases-report-artificial-intelligence-and-intellectual-property

USPTO. 2024. MPEP. 9th ed. DoC. https://www.uspto.gov/web/offices/pac/mpep/index.html

USTR. 1985. Israel Free Trade Agreement. https://ustr.gov/trade-agreements/free-trade-agreements/israel-fta

USTR. 1989. 1989 Special 301 Report. Office of the United States Trade Representative. https://ustr.gov/sites/default/files/1989%20Special%20301%20Report.pdf

USTR. 1993. The North American Free Trade Agreement: Summary of the Provisions. https://ustr.gov/sites/default/files/uploads/Countries%20Regions/africa/agreements/nafta/NAFTA%20Chapter%20Summaries.pdf

USTR. 2025. History of the United States Trade Representative. https://ustr.gov/about-us/history

Utility Dive. 2024. US data center electricity demand could double by 2030, driven by artificial intelligence: EPRI. https://www.utilitydive.com/news/artificial-intelligence-doubles-data-center-demand-2030-EPRI/717467/

Utterback, James, and William Abernathy. 1975. "A Dynamic Model of Process and Product Innovation." Omega 3(6): 639–656.

Vahdat, Amin. 2025. Ironwood: The first Google TPU for the age of inference. Google Cloud Blog. https://blog.google/products/google-cloud/ironwood-tpu-age-of-inference/

Valentino-DeVries, Jennifer, Natasha Singer, Michael Keller, and Aaron Krolik. 2018. "Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret." New York Times. https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html

Valletti, Tommaso. 2018. “How EU Markets Became More Competitive Than U.S. Markets: A Study of Institutional Drift.” NBER Working Paper No. 24700. Cambridge, MA: National Bureau of Economic Research.

Van Alstyne, Marshall, Geoffrey Parker, and Sangeet Paul Choudary. 2016. "Pipelines, Platforms, and the New Rules of Strategy." Harvard Business Review 94(4): 54-62.

Varian, Hal. 2014. "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives 28(2): 3-28.

Varian, Hal. 2007. "Position Auctions." International Journal of Industrial Organization 25(6): 1163–78.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems 30. https://papers.neurips.cc/paper/7181-attention-is-all-you-need

Vedder, Richard, and Lowell Gallaway. 1993. Out of Work. New York: Holmes and Meier.

Verizon. 2002. Verizon Communications, Inc. v. FCC. No. 00-511. US Supreme Court, May 13. https://supreme.justia.com/cases/federal/us/535/467/

Verizon. 2004. Verizon Communications, Inc. v. Law Offices of Curtis V. Trinko, LLP. No. 02-682. US Supreme Court, January 13. https://supreme.justia.com/cases/federal/us/540/398/

Vermont Yankee Nuclear Power. 1978. Vermont Yankee Nuclear Power Corp. v. Natural Resources Defense Council, Inc., et al., No. 76-419, U.S. Supreme Court, April 3. https://scholar.google.com/scholar_case?case=4079290121887020113

Vertiv. 2025. High-density cooling: A guide to advanced thermal solutions for AI and ML workloads in data centers. Linkedin Post. https://www.linkedin.com/posts/vertiv_high-density-cooling-a-guide-to-advanced-activity-7343548574292426752-5r4P

Veugelers, Reinhilde. 2018. Are European firms falling behind in the global corporate research race? Bruegel Policy Contribution 2018/06. https://www.bruegel.org/policy-brief/are-european-firms-falling-behind-global-corporate-research-race

Via Licensing Corporation. 2018. Via Expands LTE Patent Pool License Coverage to Include LTE-Advanced Pro. Press Release. www.via-corp.com/licensing/lte/index.html

Vietor, Richard. 1994. Contrived Competition. Cambridge, MA: Harvard University Press.

Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. "Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning." arXiv preprint arXiv:2211.04325. https://arxiv.org/abs/2211.04325.

Villareal, Angeles. 2017. "NAFTA and the Mexican Economy." CRS Report for Congress R42965. Washington, DC: Congressional Research Service.

Visa and Mastercard. 2001. U.S. v. Visa U.S.A. Inc. and Mastercard International, Inc., No. 98 CIV. 7076(BSJ). US District Court, S.D. New York, October 9. https://www.justice.gov/atr/case-document/file/515596/dl

Vogel, David. 1995. Trading Up. Cambridge, MA: Harvard University Press.

Von's Grocery. 1966. United States v. Von's Grocery Co., No. 303. US Supreme Court, May 31. https://supreme.justia.com/cases/federal/us/384/270/

W3Techs. n.d. “Usage Statistics and Market Share of Google Analytics for Websites.” W3Techs—Web Technology Surveys. https://w3techs.com/technologies/details/ta-googleanalytics.

Wallace, James, and Jim Erickson. 1992. Hard Drive: Bill Gates and the Making of the Microsoft Empire. New York: HarperBusiness.

WaMC. 2015. Yes, Reagan Used Trade Promotion Authority. US House of Representatives. https://waysandmeans.house.gov/2015/03/13/yes-reagan-used-trade-promotion-authority/

Ward, John William. 2017. "Digital Markets and Antitrust." Antitrust Law Journal 81(3): 919-946.

Warren, Kenneth. 2020. Administrative Law in the Political System. New York: Routledge.

Wartell, Sarah. 2009. National Economic Council. In Change for America: A Progressive Blueprint for the 44th President, edited by Mark Green and Michele Jolin. Washington, DC: Center for American Progress, 81–87.

Washburn, Brian. 2023. Omdia's new AI network traffic forecast expects surge, pressure on telcos to meet future growth. Omdia Analyst Opinion. https://omdia.tech.informa.com/om119985/omdias-new-ai-network-traffic-forecast-expects-surge-pressure-on-telcos-to-meet-future-g

Washburn, Brian. 2024. Road to 2030: AI and the Future of Network Services – Traffic Outlook and Implications. Omdia Report. https://omdia.tech.informa.com/om121972/road-to-2030-ai-and

Wassel, Bryan. 2024. “Verizon CEO says it will ‘lead AI revolution,’ improve customer service at scale.” Customer Experience Dive. https://www.customerexperiencedive.com/news/verizon-ai-revolution-customer-service-myplan-personalization/714226/

Waymo. 2025. 20 million. That’s how many times you’ve trusted Waymo to get you where you’re going. X (formerly Twitter). https://x.com/Waymo/status/2001351692240015403

Wayt, Theo. 2025. "Elon Musk's xAI Buys Building for Third Supersized Data Center." The Information. https://www.theinformation.com/articles/elon-musks-xai-buys-building-third-supersized-data-center

WCMP. 2025. “2025 Wisconsin Manufacturing Report.” https://www.wicmp.org/report-conclusions/2025-wisconsin-manufacturing-report/

Webster, Graham, Rogier Creemers, Elsa Kania, and Paul Triolo. 2017. Full Translation: China's 'New Generation Artificial Intelligence Development Plan'. DigiChina (Stanford Cyber Policy

Center). https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/

Weights and Biases. 2024. Experiment Tracking and Model Management. Weights and Biases Documentation. https://docs.wandb.ai/

Weil, Jonathan. 2025. "Is the Flurry of Circular AI Deals a Win-Win—or Sign of a Bubble?" The Wall Street Journal. https://www.wsj.com/tech/ai/is-the-flurry-of-circular-ai-deals-a-win-winor-sign-of-a-bubble-8a2d70c

Weiss, Geoff and Katherine Tangalakis-Lippert. 2024. “Trump says Google has ‘a lot of power’ and he would do ‘something’ about it — but stops short of favoring a break-up.” Business Insider. https://www.businessinsider.com/trump-google-break-up-china-comments-2024-10

Welle, Elissa. 2025. "OpenAI reportedly signs $300 billion cloud deal with Oracle." The Verge. https://www.theverge.com/ai-artificial-intelligence/776170/oracle-openai-300-billion-contract-project-stargate

Werden, Gregory. 2009. Next Steps in the Evolution of Antitrust Law: What to Expect from the Roberts Court. Journal of Competition Law and Economics 5(1): 49-74.

Werden, Gregory. 2015. Identifying Exclusionary Conduct Under Section 2: The "No Economic Sense" Test. Antitrust Law Journal 73(2) 413-433.

Werden, Gregory. 1992. “The History of Antitrust Market Delineation.” Marquette Law Review 76(1): 123–159.

Werden, Gregory. 1998. "Demand Elasticities in Antitrust Analysis." Antitrust Law Journal 66(2): 363–414.

Werden, Gregory. 2003. "The 1982 Merger Guidelines and the Ascent of the Hypothetical Monopolist Paradigm." Antitrust Law Journal 71(1): 253–75.

Wessels, Walter. 2001. The Effect of Minimum Wages on the Labor Force Participation Rates of Teenagers. Employment Policies Institute. https://epionline.org/app/studies/wessels_06-2001.pdf

West Virginia et al. 2022. West Virginia et al. v. United States Environmental Protection Agency and Michael S. Regan (Administrator of the United States Environmental Protection Agency), No. 20-1530, US Supreme Court, June 30. https://www.supremecourt.gov/opinions/21pdf/20-1530_n758.pdf

Weyerhaeuser. 2007. Weyerhaeuser Co. v. Ross-Simmons Hardwood Lumber Co., No. 05-381. US Supreme Court, February 20. https://supreme.justia.com/cases/federal/us/549/312

Weyl, Eric Glen. 2010. "A Price Theory of Multi-Sided Platforms." American Economic Review 100(4): 1642–1672.

White and Case LLP. 2025. “AI Watch: Global Regulatory Tracker – China.” White and Case Insights. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-china

White House Council of Economic Advisers. 2024. Economic Report of the President. Washington, DC: U.S. Government Publishing Office.

Wiggers, Kyle. 2024a. “Anthropic raises another $4B from Amazon, makes AWS its ‘primary’ training partner.” TechCrunch. https://techcrunch.com/2024/11/22/anthropic-raises-an-additional-4b-from-amazon-makes-aws-its-primary-cloud-partner/

Wiggers, Kyle. 2024b. “Elon Musk’s xAI Lands $6B in New Cash to Fuel AI Ambitions.” TechCrunch. https://techcrunch.com/2024/12/25/elon-musks-xai-lands-billions-in-new-cash-to-fuel-ai-ambitions/

Wikipedia. 2025a. "3G." Wikipedia. https://en.wikipedia.org/wiki/3G

Wilander, John. 2017. "Intelligent Tracking Prevention." WebKit Blog. https://webkit.org/blog/7675/intelligent-tracking-prevention/

Williams, Tom. 2025. “What We Know About OpenAI’s First Consumer Devices.” Information Age ACS. https://ia.acs.org.au/article/2025/what-we-know-about-openais-first-consumer-devices.html

Williamson, John. 1990. "What Washington Means by Policy Reform." In Latin American Adjustment. Edited by John Williamson. Washington, D.C.: Institute for International Economics, 5-20.

Williamson, Oliver. 1975. Markets and Hierarchies: Analysis and Antitrust Implications. New York: Free Press.

Williamson, Oliver. 1985. The Economic Institutions of Capitalism. New York: Free Press.

Winkler, Rolfe, and Nate Rattner. 2025. “The Driver of Apple’s Exploding Valuation Is Under Threat. See What’s at Stake.” The Wall Street Journal. https://www.wsj.com/tech/apple-stock-service-value-growth-analysis-f265a5e7

Winston, Clifford and Steven Morrison. 1997. The Fare Skies: Air Transportation and Middle America. Brookings Institution.

Wiseman, Bill, Henry Marcil, and Marc de Jong. 2025. "Semiconductors have a big opportunity—but barriers to scale remain." McKinsey & Company. https://www.mckinsey.com/industries/semiconductors/our-insights/semiconductors-have-a-big-opportunity-but-barriers-to-scale-remain

Wolla, David. 2020. Examining the “Lump of Labor” Fallacy Using a Simple Economic Model. Federal Reserve Bank of St. Louis. https://www.stlouisfed.org/publications/page-one-economics/2020/11/02/examining-the-lump-of-labor-fallacy-using-a-simple-economic-model

Wong, Kenny. 2024. What is RAG? 4 analogies for this powerful AI approach. Coda Blog. https://coda.io/blog/ai/what-is-rag-analogies

Woo, Seongjin, and Raffaele Huang. 2025. How China's DeepSeek Outsmarted America. The Wall Street Journal. https://www.wsj.com/tech/ai/china-deepseek-ai-nvidia-openai-02bdbbce?mod=hp_lead_pos1

Workday. 2023. Responsible AI practices. Workday. https://www.workday.com/en-us/artificial-intelligence/responsible-ai-practices.html

World Bank. 2023. "Charges for the Use of Intellectual Property, Receipts (BoP, current US$)" (indicator BX.GSR.ROYL.CD) and "Charges for the Use of Intellectual Property, Payments (BoP, current US$)" (indicator BM.GSR.ROYL.CD). World Development Indicators. https://data.worldbank.org/indicator/BX.GSR.ROYL.CD and https://data.worldbank.org/indicator/BM.GSR.ROYL.CD

WEF. 2025. "Towards Equitable AI: New Report Charts Path to AI Competitiveness." WEF. https://www.weforum.org/press/2025/01/towards-equitable-ai-new-report-charts-path-to-ai-competitiveness/

Wright, Joshua David. 2011. Antitrust, Multi-Dimensional Competition, and Innovation: Do We Have an Antitrust-Relevant Theory of Competition Now? In *Competition Policy and Patent Law under Uncertainty: Regulating Innovation*, edited by Geoffrey Manne and Joshua Wright. Cambridge: Cambridge University Press, 228–251.

Wright, Joshua. 2012. Defining and Measuring Search Bias: Some Preliminary Evidence. George Mason Law and Economics Research Paper No. 12-14. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2004649

Wright, Joshua. 2014. Evidence-Based Antitrust Enforcement in the Technology Sector. CPI Antitrust Chronicle 3(1): 1-9.

Wright, Joshua, and John Yun. 2018. "Ohio v. American Express: Implications for Non-Transaction Multisided Platforms." CPI Antitrust Chronicle 2(1): 1–8.

WTO. 2021. World Tariff Profiles 2021. WTO https://www.wto.org/english/docs_e/legal_e/marag_e.htm

WTO. 2008. World Trade Report 2008: Trade in a Globalizing World. WTO. https://www.wto.org/English/res_e/booksp_e/anrep_e/world_trade_report08_e.pdf

WTO. n.d.a. Overview: the TRIPS Agreement. WTO https://www.wto.org/english/tratop_e/trips_e/intel2_e.htm

WTO. n.d.b. Intellectual Property: Protection and Enforcement. WTO https://www.wto.org/english/thewto_e/whatis_e/tif_e/agrm7_e.htm

Wu, Philipp, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. 2022. "Daydreamer: World Models for Physical Robot Learning." Proceedings of the 6th Conference on Robot Learning (CoRL). https://openreview.net/forum?id=3RBY8fKjHeu.

Wu, Tim. 2018. The Curse of Bigness. New York: Columbia Global Reports.

Wu, Yonghui, et al. 2016. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.” arXiv preprint arXiv:1609.08144. https://arxiv.org/abs/1609.08144

Xu, Kevin. 2025. "DeepSeek's Three Idiosyncratic Advantages." Interconnected. https://interconnected.blog/deepseeks-three-idiosyncratic-advantages/

Yahoo Finance. 2025. AI Data Center Global Research Report 2025–2030. Yahoo Finance. https://uk.finance.yahoo.com/news/ai-data-center-global-research-091100624.html

Yale Budget Lab. 2025. "State of U.S. Tariffs: November 17, 2025." New Haven, CT: The Budget Lab at Yale. https://budgetlab.yale.edu/research/state-us-tariffs-november-17-2025.

Yan, Peifeng, Xinyuan Bai, and Zehao Shen. 2022. "Operations and Supply Chain Analysis of the Smartphone Industry: Comparing Apple and Huawei." Advances in Economics, Business and Management Research 215: 1523-1525.

Yuan, Dave, and Jair Verçosa. 2025. Does AI remove the incumbent data gravity advantage? Tidemark Capital. https://www.tidemarkcap.com/post/does-ai-remove-the-incumbent-data-gravity-advantage

Zaitsev, Peter. 2024. Vector Search in Modern Databases. Percona University (OSACon presentation). https://percona.university/wp-content/uploads/2025/03/Vector-Search-in-Modern-Databases-by-Peter-Zaitsev.pdf

Zeran. 1997. Zeran v. America Online, Inc. No. 96-433. US Court of Appeals, Fourth Circuit, November 12. https://law.justia.com/cases/federal/appellate-courts/F3/129/327

Zhu, Qiuling, Berkin Akin, H. Ekin Sumbul, Fazle Sadi, James C. Hoe, Larry Pileggi, and Franz Franchetti. 2013. A 3D-Stacked Logic-in-Memory Accelerator for Application-Specific Data Intensive Computing. Proceedings of the 2013 IEEE International 3D Systems Integration Conference (3DIC): 1–7.

Zhu, Yukun, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. “Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books.” Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015): 9–27.

Zilbershot, Dorit. 2025. Unlock Productivity: AI Agent Solutions. ServiceNow Blog. https://www.servicenow.com/blogs/2025/unlock-productivity-ai-agent-solutions

Zittrain, Jonathan. 2008. The Future of the Internet—And How to Stop It. New Haven, CT: Yale University Press.

Zuboff, Shoshana. 2019. The Age of Surveillance Capitalism. New York: PublicAffairs.

Zubrow, Keith. 2021. “Facebook Whistleblower Says Company Incentivizes ‘Angry, Polarizing, Divisive Content.’” CBS News. https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-60-minutes-polarizing-divisive-content/

Zuckerberg, Mark. 2024. "Open Source AI Is the Path Forward." Meta Newsroom. https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/

Zulli, Diana, and David James Zulli. 2022. "Extending the Internet Meme." New Media and Society 24(8): 1872–1890.