AI Has No Rights: from System-Based to Stakeholder-Based AI Governance

Damian Hodel
University of Washington
Seattle, USA
hodeld@uw.edu

Lindah Kotut
University of Washington
Seattle, USA
kotut@uw.edu

ABSTRACT

Prevailing approaches to govern artificial intelligence (AI) focus on dictating how AI should behave, such as ensuring it avoids social biases and the spread of misinformation. While these system-based frameworks provide reasonable guidance, they risk promoting AI that largely benefits AI operators and dominant social groups, while further harming already marginalized individuals. In this position paper, we argue for a shift from AI safety—how AI should behave—to human safety: individuals should only be impacted by AI systems they explicitly approve. First, we explore the challenges inherent in system-based AI governance including the lack of operationalizable definitions of AI and its desired behavior, the failure to address the collective impact of AI systems, and the situated harm of impacted stakeholders. Second, we propose an approach for enforcing our new ideal. We recommend that AI operators (1) clearly specify each system's purpose and alignment approach, which is necessary to (2) secure informed approval from stakeholders before the system's impact. To address the challenges of the approval process, we propose a complaint mechanism that allows users and indirect stakeholders to report systems they did not approve or that deviate from their approved impact.

KEYWORDS

AI Governance, AI Alignment, System Pluralism, Pluralistic Alignment, Artificial Intelligence

ACM Reference Format:
Damian Hodel and Lindah Kotut. 2025. AI Has No Rights: from System-Based to Stakeholder-Based AI Governance. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND BACKGROUND

AI increasingly interferes in our lives, creating beneficial but also harmful experiences depending on an individual's social, geographical, and historical context [4, 9, 40]. The highest risks stem from so-called general-purpose AI—systems with broad capabilities designed to be used for everything and to impact everyone. Mitigating the wide spectrum of reported harm requires effective regulation. Current approaches, such as the European Union AI Act (EU AI Act) [31] and the non-binding National Institute of Standards and Technology's "AI Risk Management Framework" (NIST framework) [30], are system-based policies aiming to regulate AI systems rather than their use cases [21]. While these frameworks provide reasonable standards, such as prohibiting the exhibition of social biases, system-based governance presents significant challenges. First, to effectively regulate AI, there would need to be a universal and operationalizable definition of an ideal AI system—which does not currently exist. Second, system-based regulation limits control over the collective impact of multiple (homogeneous) AI systems, which are responsible for many forms of harm and performance issues [18, 24, 25]. Third, system-based regulation struggles to account for situated harm and located accountability in a complex AI landscape, as well as the unforeseen behavior and use cases of AI systems.

Unlike humans, AI systems do not hold any rights. AI governance must solely ensure that AI systems serve both their directly and indirectly impacted stakeholders[1]. To model this idea, we propose a new *ideal*: **individuals should only be impacted by AI systems they explicitly approve**, along with the systems' specific and collective pros and cons. The work presented here results from a research project, where we used a speculative approach to assess value tensions among diverse (fictionalized) AI creators.

The remainder of this paper is as follows: First, we explore the problem space of system-based regulatory approaches to motivate our proposed shift in guiding AI governance. Second, to make this approach tangible for discussion, we outline a potential enforcement of the proposed ideal based on an informed approval process and a complaint mechanism. Third, we illustrate our proposed implementation approach with a specific scenario grounded in current AI applications: AI in classrooms.

Ambiguous definitions. To design effective system-level regulation, a universal understanding of an ideal system and operationalizable definitions of both desired and undesired behavior are required. However, such definitions do not exist, whether for human values like justice and fairness [13, 42], or for potential guardrails like misinformation [35, 37] and social biases [19]. This ambiguity in definitions creates loopholes that AI operators can intentionally or unintentionally exploit to build systems that serve their interests rather than those of the impacted individuals [15, 20, 39].

Collective impact. Focusing on systems rather than AI use overlooks the many issues arising from AI monoculture [18, 25]. Prior work shows that multiple homogeneous systems can collectively create outcomes that are not only unfair for specific social groups but also worse overall across all stakeholder groups [6, 24, 26].

[1] Stakeholders are "those who are or will be significantly implicated by the technology" [14, p. 35], whereas indirect stakeholders refer to individuals who are affected indirectly through others' use of AI systems [4, 14].

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Futhermore, while a single system may superficially accommodate diverse perspectives [23, 34], it can only accommodate one of many competing theories on how such aggregation of perspectives should be executed [26]. These issues of AI monoculture can be addressed through system pluralism—an AI landscape consisting of diverse systems.

Situated harm and nested accountability. Whether a specific behavior, such as a piece of text or an image, is harmful depends on one's social, geographical, and historical situation [22, 40]. However, a policy design that evaluates systems rather than applications can hardly account for the context-sensitive nature of harm, nor does it hold humans involved along the value chain accountable for their contributions to potential harmful outcomes [2, 41]. Consider this example: Company A offers a general-purpose generative language model, which Company B fine-tunes for educational purposes. Suppose a teacher uses this fine-tuned system to evaluate students' written essays, resulting in unfair grading for some students because the language model undervalues indigenous knowledge. An effective policy would steer every individual involved in the AI process to make decisions that prevent this situated harm. These challenges are further amplified by technical issues that limit reliable control over AI system behavior.

(Regulating) AI systems inherently involve a trade-off between benefits and harms. The net impact of this trade-off varies from case to case and individual to individual. Therefore, the overall net loss across the AI landscape and among stakeholders can be minimized when stakeholders have the ability to decide on a case-by-case basis whether the specific benefits outweigh the risks of harm. AI governance should reflect this idea.

2 APPROVAL-BASED AI GOVERNANCE

In this section, we present a potential implementation of our proposed ideal. The premise of our approach is straightforward: before impacting humans (through the development or deployment of a system), AI operators must obtain stakeholders' informed approval. To address the challenges of the approval process, we propose a complaint mechanism (see Figure 1).

Informed and compensated approval. To make AI operators accountable for their impact, we propose a two-step procedure comprising an alignment statement and an approval statement (see Figure 1).

The goal of the **alignment statement** is to provide sufficient information for stakeholders to make informed decisions about whether to approve the given system's impact. Specifically, the alignment statement should address questions related to the system's purpose, data, energy consumption, alignment approach, its underlying normative framework, fine-tuning, human involvement, the anticipated impacted stakeholders and an assessment of the risks and benefits for them. It can build on prior work and established practices such as data statements [3], datasheets [17], model cards [28], and impact assessments [33].

The **approval statement** serves as the formal endorsement, addressing key considerations such as: the individuals granting approval, the representation of social groups in the case of large

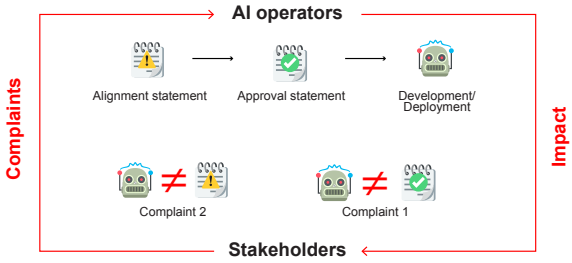

Figure 1: Approval-based AI governance. Before impacting individuals through development and deployment, AI operators must (1) state the purpose of their system along its alignment approach, and (2) obtain approval from impacted stakeholders. Stakeholders can file complaints if (1) they are impacted by an AI system they did not approve, or (2) they are impacted by an AI system in a manner inconsistent with their approval.

populations, strategies to prevent tokenizing certain groups [16], the information shared, compensation for participants that reflects the effort required to understand the AI system, actions to be taken if the system's behavior deviates from the approval statement, and the validity period of the approval. To address the limited accommodation of diverse perspectives within a single system, approval can be sought also for a combination of diverse systems [26]. The idea of stakeholder approval builds on prior concepts related to agreement-based approaches [e.g., 8, 43], while the approval process can draw from stakeholder-based alignment approaches [e.g., 5, 7, 10, 14]. Notably, this approval process differs from user agreements, as it also includes indirect stakeholders who do not have the same conflicts of interest as direct users [cf. 38]. A derived approval score—the ratio of stakeholders who approve a system to the total number impacted by it—could be used to evaluate AI use cases.

Given the varying forms of AI and its impacts, the information and steps required for informed approval highly depend on the purpose of the AI system. For example, the effort required for a system restricted to a small company's internal use is significantly smaller than for a general-purpose system like ChatGPT. Because our approach relies on stakeholder approval of specific AI use cases, rather than regulating AI in general [21], a regulatory framework can allow AI operators some flexibility in tailoring their alignment and approval statements to the specific AI use case.

Complaint mechanism. Over-reliance on self-assessment offers loopholes for AI operators [39], such as tokenizing social groups [16], underestimating the risks of AI systems [39], or undervaluing minority perspectives [5]. To address those, we propose a complaint mechanism. The premise is that individuals who can demonstrate they are directly or indirectly impacted by an AI system without having approved this impact should have the ability to file complaints against the operator. For legal implementation, this mechanism can build on frameworks like the EU AI Act, which allows complaints if a system infringes regulations [31, 39]. Broadly, operators face two primary types of complaints: (1) Stakeholders are impacted by a system they did not approve and (2) Stakeholders

are impacted by an AI system in a manner inconsistent with their approval (see Figure 1).

This regulation compels AI operators to design systems that align with stakeholders' interests and empowers them to determine when the benefits of a specific AI use outweigh its potential harms. At the level of the AI landscape, this approach leads toward system pluralism, addressing the issues resulting from AI monoculture.

Legal consequences of complaints—such as financial penalties, halts in development or deployment, or required system adjustments—should be reasonable and aligned with the goal of preventing harm and ensuring an equitable distribution of benefits. A mere political or aesthetic disagreement with an AI system may not suffice for a complaint. However, if stakeholders can demonstrate, for instance, that a language model systematically undervalues or ignores indigenous knowledge or minoritized dialects, thereby perpetuating epistemic injustice [12, 22], a complaint from impacted stakeholders would be well justified.

If stakeholders do not approve of a system or file a complaint, AI operators can take one of three actionable adjustments: (1) revise the alignment approach to produce outcomes more aligned with the interests of the stakeholders who approved the system, (2) modify the AI value chain or application to affect fewer or different stakeholders, or (3) seek approval for a combination of diverse systems that better reflect the diversity of impacted stakeholders. For instance, the operator of a general-purpose AI could develop additional systems providing different behaviors [11, 32, 44] or collaborate with independent organizations to integrate systems developed with varying perspectives [26]. This approach calls for new interactive designs that allow users to choose from the different integrated systems.

We define an *AI operator* as the legal entity responsible for providing and/or operating an AI system for a specific application. An *AI system* encompasses the entire value chain up to the given application as reflected in alignment and approval statement. Since a given AI application can be built on another AI system further down the value chain, individuals or organizations using a system—thereby impacting others—can act as both operators of the AI system and stakeholders in the system on which their application is based. To hold responsible actors accountable, the complaint mechanism could follow a cascade model along the value chain. Unlike system-based approaches, our proposed model allows for addressing the complexities of AI applications on a case-by-case basis. We illustrate such a case in the example of *AI in classrooms* in the next section.

3 AI IN CLASSROOMS SCENARIO

A high school teacher wants to use a language model like ChatGPT as an assistant, for tasks such as creating quizzes, grading essays, and helping students look up information during class. In this scenario, the stakeholders of the AI system include the students (legally represented by their parents) and the school administrators, who are responsible for upholding educational goals. In the alignment statement, the teacher describes how the language model is expected to behave and how it will be used. Additionally, it includes details on the duration of approval (e.g., one school year) and specifies the measures to be taken (e.g., the teacher refrains from using the language model) if any information in the alignment statement is later found to be false. Based on the provided information, stakeholders then decide whether to approve the alignment statement.

There are numerous ethical concerns associated with the use of such a system in education, including potential privacy violations, social biases, and the loss of autonomy for both teachers and students in cases of automated assessment [1]. The proposed policy ensures that these case-specific ethical issues are addressed. For instance, school administrators might worry that students could anthropomorphize the language model. They could approve its use on the condition that the AI does not respond in human-like language. Similarly, parents might have specific concerns, such as the system's environmental impact, potential copyright issues, or whether it exhibits toxic language. Before granting approval, parents could request this information from the teacher, who would share the details provided in the alignment statement received from the language model provider when they approved it as stakeholders. Here, the teacher is both the AI operator of the classroom AI and a stakeholder of the underlying language model. If parents approve the use of the classroom AI but later discover that the information regarding its environmental impact was incorrect, the teacher, as the operator, would cease using the AI in the classroom. Additionally, in the role of a stakeholder, the teacher could file a complaint against the language model provider. This illustrates how the proposed approach can follow a cascade model to enforce located accountability [36, 40] in the complex AI landscape. It is also conceivable that some parents might disapprove of a system because of its sociopolitical values [27, 29]. Acknowledging that AI is not inevitable (in classrooms), there are two alternatives for the teacher and school: either refrain from using the language model or include a second language model from a different AI provider that reflect diverse value sets.

4 CONCLUSION

System-based AI governance can offer valuable guidance, but it may not be enough to ensure the equitable distribution of benefits among impacted stakeholders. To address this, we propose a new approval-based framework and outline a potential implementation strategy based on an alignment statement, an approval statement, and a complaint mechanism.

The approach proposed here is not a fully developed regulatory framework but rather a conceptual contribution aimed at rethinking and discussing how AI can be regulated. The idea that individuals should only be impacted by AI systems they explicitly approve is practically infeasible. Much like the notion that one system can fit everyone. Therefore, we contrast our approach with the existing one of defining and regulating systems intended to serve everyone and everything, which, in reality, results in AI serving particular people, species, and places, thereby creating social disparities. Overall, we anticipate that our approach will help mitigate these disparities.

REFERENCES

[1] Selin Akgun and Christine Greenhow. 2022. Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics* 2, 3 (Aug. 2022), 431–440. https://doi.org/10.1007/s43681-021-00096-7

[2] Blair Attard-Frost and David Gray Widder. 2024. The Ethics of AI Value Chains. (Sept. 2024).

[3] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[5] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports* 14, 1 (March 2024), 6616. https://doi.org/10.1038/s41598-024-56648-4

[6] Emily Black, John Logan Koepke, Pauline T. Kim, Solon Barocas, and Mingwei Hsu. 2024. Less discriminatory algorithms. *Geo. LJ* 113 (2024), 53. Publisher: HeinOnline.

[7] Alessandra Calvi and Dimitris Kotzinos. 2023. Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1229–1245. https://doi.org/10.1145/3593013.3594076

[8] Georgina Curto and Flavio Comim. 2023. SAF: Stakeholders' Agreement on Fairness in the Practice of Machine Learning Development. *Science and Engineering Ethics* 29, 4 (Aug. 2023), 29. https://doi.org/10.1007/s11948-023-00448-y

[9] AI Incident Database. 2024. Welcome to the Artificial Intelligence Incident Database. https://incidentdatabase.ai/

[10] Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Montr\'{e}al QC Canada, 38–48. https://doi.org/10.1145/3600211.3604661

[11] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. http://arxiv.org/abs/2406.15951 arXiv:2406.15951 [cs].

[12] Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Clarendon Press. Google-Books-ID: lncSDAAAQBAJ.

[13] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (April 2021), 136–143. https://doi.org/10.1145/3433949

[14] Batya Friedman and David G Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

[15] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024. The Ethics of Advanced AI Assistants. https://doi.org/10.48550/arXiv.2404.16244 arXiv:2404.16244 [cs].

[16] Timnit Gebru and Remi Denton. 2024. Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and Fostering Community-Rooted Computer Vision Research. *Foundations and Trends® in Computer Graphics and Vision* 16, 3 (2024), 215–321. https://doi.org/10.1561/0600000102

[17] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. https://doi.org/10.1145/3458723

[18] Timnit Gebru and Émile P. Torres. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday* (April 2024). https://doi.org/10.5210/fm.v29i4.13636

[19] Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models. http://arxiv.org/abs/2305.12757 arXiv:2305.12757 [cs].

[20] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. Unsolved Problems in ML Safety. http://arxiv.org/abs/2109.13916 arXiv:2109.13916 [cs].

[21] Emmie Hine and Luciano Floridi. 2023. The Blueprint for an AI Bill of Rights: In Search of Enaction, at Risk of Inaction. *Minds and Machines* 33, 2 (June 2023), 285–292. https://doi.org/10.1007/s11023-023-09625-1

[22] Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. 2024. Epistemic Injustice in Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 684–697. https://ojs.aaai.org/index.php/AIES/article/view/31671

[23] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. http://arxiv.org/abs/2404.16019 arXiv:2404.16019 [cs].

[24] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (June 2021), e2018340118. https://doi.org/10.1073/pnas.2018340118

[25] Seth Lazar and Alondra Nelson. 2023. AI safety on whose terms? *Science* 381, 6654 (July 2023), 138–138. https://doi.org/10.1126/science.adi8982

[26] Christina Lu and Max Van Kleek. 2024. Model Plurality: A Taxonomy for Pluralistic AI. (Oct. 2024).

[27] John Levi Martin. 2023. The Ethico-Political Universe of ChatGPT. *Journal of Social Computing* 4, 1 (March 2023), 1–11. https://doi.org/10.23919/JSC.2023.0003 Conference Name: Journal of Social Computing.

[28] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[29] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring ChatGPT political bias. *Public Choice* (Aug. 2023). https://doi.org/10.1007/s11127-023-01097-2

[30] National Institute of Standards and Technology (US). 2024. *Artificial intelligence risk management framework : generative artificial intelligence profile*. Technical Report error: 600-1. National Institute of Standards and Technology (U.S.), Gaithersburg, MD. error: 600–1 pages. https://doi.org/10.6028/NIST.AI.600-1

[31] EUROPEAN PARLIAMENT. 2024. REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[32] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 71095–71134. https://proceedings.neurips.cc/paper_files/paper/2023/file/e12a3b98b67e8395f639fde4c2b03168-Paper-Conference.pdf

[33] Dillon Reisman, Jason Schultz, Kate Crawford, and Whitaker Meredith. 2018. Algorithmic Impact Assessment: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY. *Impact Assessment* 13, 1 (April 2018), 3–30.

[34] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. (Feb. 2024). https://arxiv.org/abs/2402.05070

[35] Brian G. Southwell, J. Scott Babwah Brennen, Ryan Paquin, Vanessa Boudewyns, and Jing Zeng. 2022. Defining and Measuring Scientific Misinformation. *The ANNALS of the American Academy of Political and Social Science* 700, 1 (March 2022), 98–111. https://doi.org/10.1177/00027162221084709 Publisher: SAGE Publications Inc.

[36] Lucy Suchman. 2002. Located accountabilities in technology production. 14 (2002).

[37] Joseph Uscinski, Shane Littrell, and Casey Klofstad. 2024. The Importance of Epistemology for the Study of Misinformation. *Current Opinion in Psychology* (Jan. 2024), 101789. https://doi.org/10.1016/j.copsyc.2024.101789

[38] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. Association for Computing Machinery, New York, NY, USA, 973–990. https://doi.org/10.1145/3319535.3354212

[39] Sandra Wachter. 2024. Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond. *SSRN Electronic Journal* (2024). https://doi.org/10.2139/ssrn.4924553

[40] David Gray Widder. 2024. Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1295–1304. https://doi.org/10.1145/3630106.3658973

[41] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society* 10, 1 (Jan. 2023), 20539517231177620. https://doi.org/10.1177/20539517231177620 Publisher: SAGE Publications Ltd.

[42] Stephen Tze-Inn Wu, Daniel Demetriou, and Rudwan Ali Husain. 2023. Honor Ethics: The Challenge of Globalizing Value Alignment in AI. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 593–602. https://doi.org/10.1145/3593013.3594026

[43] Tan Zhi-Xuan. [n. d.]. What Should AI Owe To Us? Accountable and Aligned AI Systems via Contractualist AI Alignment. ([n. d.]). https://www.alignmentfor

um.org/posts/Cty2rSMut483QgBQ2

[44] Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 10586–10613. https://doi.org/10.18653/v1/2024.findings-acl.630