Computational Foundations of Human Social Intelligence

by

Max Kleiman-Weiner

B.S., Stanford University (2009)
MSc, University of Oxford (2010)
MSc, University of Oxford (2012)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author . . . .
Department of Brain and Cognitive Sciences
May 4, 2018

Certified by . . . .
Joshua B. Tenenbaum
Professor of Computational Cognitive Science
Thesis Supervisor

Accepted by. . . . .
Matthew A. Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

Computational Foundations of Human Social Intelligence

by

Max Kleiman-Weiner

Submitted to the Department of Brain and Cognitive Sciences
on May 4, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis develops formal computational cognitive models of the social intelligence underlying human cooperation and morality. Human social intelligence is uniquely powerful. We collaborate with others to accomplish together what none of us could do on our own; we share the benefits of collaboration fairly and trust others to do the same. Even young children work and play collaboratively, guided by normative principles, and with a sophistication unparalleled in other animal species. Here, I seek to understand these everyday feats of social intelligence in computational terms. What are the cognitive representations and processes that underlie these abilities and what are their origins? How can we apply these cognitive principles to build machines that have the capacity to understand, learn from, and cooperate with people?

The overarching formal framework of this thesis is the integration of individually rational, hierarchical Bayesian models of learning, together with socially rational multi-agent and game-theoretic models of cooperation. I use this framework to probe cognitive questions across three time-scales: evolutionary, developmental, and in the moment. First, I investigate the evolutionary origins of the cognitive structures that enable cooperation and support social learning. I then describe how these structures are used to learn social and moral knowledge rapidly during development, leading to the accumulation of knowledge over generations. Finally I show how this knowledge is used and generalized in the moment, across an infinitude of possible situations.

This framework is applied to a variety of cognitively challenging social inferences: determining the intentions of others, distinguishing who is friend or foe, and inferring the reputation of others all from just a single observation of behavior. It also answers how these inferences enable fair and reciprocal cooperation, the computation of moral permissibility, and moral learning. This framework predicts and explains human judgment and behavior measured in large-scale multi-person experiments. Together, these results shine light on how the scale and scope of human social behavior is ultimately grounded in the sophistication of our social intelligence.

Thesis Supervisor: Joshua B. Tenenbaum
Title: Professor of Computational Cognitive Science

Acknowledgments

This thesis is aimed at understanding human social intelligence and it wouldn't have been possible without some incredibly intelligent and cooperative humans.

Josh Tenenbaum, you have shaped each and every chapter of this thesis. It has been an honor to work with and learn from someone as passionate, curious, and brilliant as you are. Thank you for letting me share in your intellectual vision. I hope this is just the beginning.

I am grateful to have been mentored by such a wonderful thesis committee: Rebecca Saxe, you've shown me what real scientific thinking looks like. Thank you for always challenging me to go deeper and helping me get there. Drazen Prelec, your wide breadth of knowledge has opened so many doors. Thank you for pushing me to work across disciplines. Fiery Cushman, you are a fountain of creativity. Thank you for helping me distill and polish each idea into its pithiest bit.

The fun part of this research has been primarily due to a fantastic group of collaborators and mentors: Kelsey Allen, Joe Austerweil, Edmond Awad, Jack Cao, Nick Chater, Joe Halpern, Mark Ho, Dae Houlihan, Richard Kim, Julia Leonard, Sydney Levine, Michael Littman, Iyad Rahwan, David Rand, and Alex Shaw.

My fellow members of CoCoSci and friends in Cambridge, thank you for a half decade of stimulating discussions and fun: Chris Baker, Justin Brooke, Noam Brown, Andres Campero, Eyal Dechter, Johnny and Carolyn DuPont, Sam Gershman, Kevin Ellis, Nathalie Fernandez, Michael Janner, Julian Jara-Ettinger, Boris Hanin, Josh Hartshorne, Luke Hewitt, Natasha Jaques, Alex Kell, Eliza Kosoy, Peter Krafft, Tejas Kulkarni, Brenden Lake, Kyle Mahowald, Jon Malmoud, John McCoy, Max Nye, Jonathan Phillips, Dave Reshef, Eric Schulz, Max Siegal, Kevin Smith, Amir Soltani, Felix Sosa, Tomer Ullman, Ilker Yildirim, Will Whitney, Elizabeth Wu, and Yibiao Zhao. John Frank, for teaching me street smarts. Tobi Gerstenberg, for late night conferences and hands-on mentorship in both surfing and science. Tejas Kulkarni, for endless optimism in the face of the near impossible. Dan Roberts, for a decade of inside jokes, deep philosophy, and other science. Pedro Tsividis, for all manners of questing.

Thank you to the brilliant, creative, and incredibly hard working students I've worked with: Alyssa Dayan, Michael Shum, Alejandro Vientós, Lily Zhang, and Penghui Zhou.

I'm deeply indebted to the Hertz Foundation for funding both my graduate education and introducing me to Ashvin Bashyam, Jesse Engreitz, Kelly Moynihan, Josh Mueller, Vyas Ramanan, Thomas Segall-Shapiro, DJ Strouse, and Jim Valcourt. I also thank the NSF Graduate Research Fellowship and the Center for Brain, Minds, and Machines (CBMM) for funding.

Finally, my family Rose, Mom, and Dad. I love you so much. Thank you for supporting me in anything and everything. I'm so lucky to have you. Even after all these years I feel I still have so much to learn from you.

Contents

List of Figures

Chapter 1

Introduction

Ten thousand years ago, *Homo sapiens* living as hunters and gatherers began developing the first notions of what eventually would become agriculture and industry. Today, we have cities, airplanes, computers, medicines, science, and supermarkets. How did we get so much from so little so quickly? To summarize a few hundred millennia of human history in a few sentences: we collectively accumulated knowledge and technology by sharing what was learned with each succeeding generation (Boyd & Richerson, 1988; Deutsch, 2011; Henrich, 2015). This feat is driven by the power of our social intelligence. We collaborate with others to accomplish together what none of us could do on our own, share the benefits of collaboration fairly, and trust others to do the same (Humphrey, 1976; Tomasello, 1999, 2014). Even young children work and play collaboratively guided by normative principles, with a sophistication unparalleled in other animal species (Vygotsky, 1978; Warneken & Tomasello, 2006; Herrmann, Call, Hernández-Lloreda, Hare, & Tomasello, 2007; Spelke & Kinzler, 2007; Hamlin, 2013).

These successes aren't limited to the economic or scientific. We also develop sophisticated abstract entities such as political systems and organized institutions that can enhance our collaborations and amplify our knowledge (Coase, 1960; Posner, 1973; H. P. Young, 2001; Friedman, 2001). Moral and ethical systems have on average accumulated rights and rules that have trended towards greater equality, more freedom, and less conflict (Pinker, 2011). Even on a micro-scale these systems permeate human life. Hunters and gatherers shared food according to complex rituals and rules and today we decide whose name goes where on a scientific manuscript according to norms no less arcane.

Yet even with these successes, cooperation is anything but inevitable. A well studied challenge is the problem of conflicting incentives: cooperation requires individuals to bare personal costs in order to create these collective benefits which can lead to a "tragedy of the commons" (Hardin, 1968). Successful cooperation also poses hard cognitive challenges (Cosmides & Tooby, 1992; Pinker, 1997). How to distinguish friend from foe? Who should we learn moral principles from and how do we learn them so quickly? When is someone's action deserving of condemnation or praise? What are reputations, how do we learn them, and when do we manage our own? Compared to the variety and complexity of these decisions and judgments, our experiences are sparse. We rarely encounter the same exact situation twice. Yet we solve these problems everyday, whether its our first day of elementary school or out to dinner as part of a job interview. In the natural world, human social cognition is the most sophisticated known solution to these problems. In contrast, our best artificial intelligences are often exceeded by the social skills of even a kindergartner. How do we learn so much from so little so quickly?

Economists and computer scientists have developed formal quantitative frameworks to try to understand these abilities, game theory being a prominent example (Binmore, 1994; Gintis, 2009). However, these frameworks do not capture some of the most interesting aspects of human cooperation. Compared to behavioral automata (such as Tit-For-Tat) that are hand-designed for cooperation in a single task, or reinforcement learning algorithms that require long periods of trial-and-error learning, people cooperate much more flexibly with much less experience (Fudenberg & Levine, 1998; Sigmund, 2010). In real life (unlike a repeated prisoners dilemma), each social interaction is unique and complex. Real world cooperation requires coordination over extended actions that unfold in space and time, as well as the ability to plan in an infinite range of novel environments with potentially uncertain and unequal payoffs. Distinctively human cooperation also requires abstraction: we learn and plan with abstract moral principles that determine how the benefits of cooperation should be distributed and how those who fail to cooperate should be treated. In contrast to existing formal frameworks, psychologists have identified rich cognitive capacities such as "theory of mind," "joint intentions," or "moral grammar" that might underlie human cooperation (Wellman, 1992; Tomasello, Carpenter, Call, Behne, & Moll, 2005; Mikhail, 2007). But without quantitative precision, their theories leave open many different interpretations and often fail to generate definite, testable predictions or explanations that could satisfy an economist or computer scientist.

In this thesis I aim to combine the best features of these different disciplines by reverse-engineering the cognitive capacities of social intelligence that psychologists have proposed. I do so in terms sufficiently precise and rigorous that we can understand the functional role of these capacities as an engineer would (Marr, 1982; Pinker, 1997; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). That is, I aim to explain how our social intelligence works by asking what cognitive principles will be needed to recreate it in machines. The specific tools I use integrate Bayesian models of learning and multi-agent planning algorithms from artificial intelligence together with analytical frameworks from game theory and evolutionary dynamics. These models are both formally precise and make possible fine-grained quantitative predictions about complex human behavior in diverse domains. I test these predictions in large-scale multi-person experiments.

As philosophers going back to Hume have noted, “there can be no image of virtue, no taste of goodness, and no smell of evil” (Hume, 1738; Prinz, 2007). How then can we learn concepts like moral theories when there is no explicitly moral information in our perceptual input? If human cooperation builds on moral and social concepts that are richer than the relative poverty of the stimulus, then something else inside the mind must make up the difference.

My thesis proposes that the human mind bridges this gap by recursively representing mental models of other agents that have motivations and minds of their own (Dennett, 1989). These representations allow us to “read the minds” of other people by recovering the latent causal factors such as the intentions, beliefs, and desires that drove the agent to act (Heider, 1958; Wellman, 1992; C. L. Baker, Saxe, & Tenenbaum, 2009; C. L. Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). They also allow us to predict what an agent is likely to do next through forward simulation, or even consider, counterfactually, what an agent would have done differently had circumstances been different.

I use the computational structure of these abstract representations to study how they enable flexible social intelligence across three time-scales: evolutionary, developmental, and in the moment. What are the *evolutionary* origins (biological or cultural) of our moral and social knowledge and how do they enable distinctively human cooperation? How is this knowledge rapidly learned with high fidelity during *development*, accumulating over generations and giving rise to cumulative cultural? Finally, how is social and moral knowledge generalized and deployed *in the moment*, across an infinitude of possible situations and people, and how is this knowledge collectively created? To answer these questions, I investigate the cognitive structures that span across these time-scales: they emerge from evolution out of a world of non-social agents, support acquisition during development, and enable flexible reasoning and planning in any particular situation.

1.1 Evolutionary Origins

> *Natural selection does not forbid cooperation and generosity; it just makes them difficult engineering problems.*
>
> – Steve Pinker, How the Mind Works

Explaining the evolution of cooperation – where individuals pay costs to benefit others – has been a central focus of research across the natural and social sciences for decades (Hardin, 1968; Ostron, 1990; Axelrod, 1985; M. A. Nowak, 2006; Rand & Nowak, 2013). A key conclusion that has emerged from this work is the centrality of reciprocity: evolutionary game theoretic models have robustly demonstrated how repeated interactions between individuals (direct reciprocity) and within groups (indirect reciprocity) can facilitate the evolutionary success of cooperation. Although these models can provide fundamental insights due to their simplicity, this simplicity also imposes stark limits.

In particular, the winning cooperative strategies identified by these models, such as tit-for-tat (M. A. Nowak & Sigmund, 1992) or win-stay-lose-shift (M. Nowak & Sigmund, 1993), can rarely be applied to actual human interactions. This is because these strategies are defined within the context of one specific game (typically a particular Prisoner's

Dilemma). If confronted with an even slightly different game representing a slightly different decision, nothing that agents in a typical evolutionary simulation have learned generalizes at all. For example, agents who cooperate in a prisoner's dilemma – that is, to choose the C row or column in a 2×2 (or $[C,D] \times [C,D]$) matrix – haven't learned to be altruistic in dictator games or to be trusting in public goods games, even though these are all very similar. This is because what these automata have learned is just a policy of how to act in a particular setting without any abstract knowledge of reciprocity.

Human interactions, in contrast, are almost infinitely varied. Even when the same two people interact in the same context, no two interactions have exactly the same payoff structure; and, more broadly, we engage in all manner of different interactions where the number of participants, the options available to each participant, and the resulting payoffs differ markedly (and often unpredictably). Because of this variation, it is implausible (and impractical) to imagine that people learn a specific strategy for every possible game. Rather than a specific strategy specifying how to play a specific game, humans need a general strategy which can be applied to cooperate across contexts. That is, sophisticated cooperators need an abstract theory of reciprocity.

In **Chapter 2** I introduce a new approach to the evolution of cooperation which solves this challenge. I do so by leveraging the key insight that people use others' actions to make inferences about their beliefs, intentions, and desires (i.e. humans have theory of mind). This stands in marked contrast to the standard game theoretic strategies, which respond only to other agents' actions, without making inferences about why a given agent chose a given action. Instead endowing agents with theory of mind allows them to have a general utility function which they can apply across all possible interactions. I show that such a strategy – specifically, a conditional cooperator that uses Bayesian inference to preferentially cooperate with others who have the same strategy – enables the evolution of cooperation in a world where every interaction is unique. Furthermore, even in the context of repeated play of one specific iterated Prisoner's Dilemma, natural selection favors our cognitively endowed strategy over all of the standard behavioral strategies even in specific contexts those strategies were designed for. And finally, the framework seamlessly integrates direct and indirect reciprocity, with our cognitively endowed agent leading to the evolution of cooperation when pairs of players interact repeatedly, when pairs play one-shot games that are observed by others, or any combination of the two. Thus, I show that cognitive complexity enables the evolution of cooperation more effectively than purely behaviorist strategies explaining in part the scale and scope of human cooperation. These results are also suggestive of how the challenge of cooperation can drive the evolution of cognitive complexity – a defining feature of humankind.

1.2 Learning and Development

> *A recipe book written for great chefs might include the phrase "poach the fish in a suitable wine until almost done," but an algorithm for the same process might begin "choose a white wine that says 'dry' on the label; take a corkscrew and open the bottle; pour an inch of wine in the bottom of a pan; turn the burner under the pan on high;..." – a tedious breakdown of the process into dead-simple steps, requiring no wise decisions or delicate judgments or intuitions on the part of the recipe-reader.*
>
> – Daniel Dennett, Intuition Pumps And Other Tools for Thinking

> *That which is hateful to you, do not do to your neighbor. That is the whole Torah; the rest is the explanation.*
>
> – Hillel the Elder

Scaling cooperation across the full range of social life confronts us with the need to tradeoff the interests and welfare of different people: between our own interests and those of others, between our friends, family or group members versus the larger society, people we know who have been good to us or good to others, and people we have never met before or never will meet. These trade-offs encoded as a system of values are basic to any commonsense notion of human morality. While some societies view preferential treatment of kin as a kind of corruption (nepotism), others view it as a moral obligation (what kind of monster hires a stranger instead of his own brother?). Large differences both between and within cultures pose a key learning challenge: how to infer and acquire appropriate values, for moral trade-offs of this kind? Can we build moral machines that learn human values like young children and apply them to novel situations?

In **Chapter 3** I develop a computational framework for understanding the structure and dynamics of moral learning, with a focus on how people learn to trade off the interests and welfare of different individuals in their social groups and the larger society. We posit a minimal set of cognitive capacities that together can solve this learning problem: (1) an abstract and recursive utility calculus to quantitatively represent welfare trade-offs; (2) hierarchical Bayesian inference to understand the actions and judgments of others; and (3) meta-values for learning by value alignment both externally to the values of others and internally to make moral theories consistent with one's own attachments and feelings. Our model explains how children can build from sparse noisy observations of how a small set of individuals make moral decisions to a broad moral competence, able to support an infinite range of judgments and decisions that generalizes even to people they have never met and situations they have not been in or observed. It also provides insight into the causes and dynamics of moral change across time, including cases when moral change can be rapidly progressive, changing values significantly in just a few generations, and cases when it is likely to move more slowly.

1.3 Planning and Reasoning In-the-Moment

> *Any social transaction is by its a nature a developing process and the development is bound to have a degree of indeterminacy to it. Neither of the social agents involved in the transaction can be certain of the future behavior of the others; as in Alice's game of croquet with the Queen of Hearts, both balls and hoops are always on the move. Someone embarking on such a transaction must therefore be prepared for the problem itself to alter as a consequence of his attempt to solve it – in the very act of interpreting the social world he changes it. Like Alice he may well be tempted to complain "You've no idea how confusing it is, all the things being alive"; that is not the way the game*

is played at Hurlingham – and that is not the way that non-social material typically typically behaves.

– Nicholas Humphrey, The Social Function of Intellect

Think how hard physics would be if particles could think.

– Murray Gell-Mann

To reverse-engineer human cooperation, we need new tasks that highlight the flexibility of human cognition. Inspired by stochastic games studied in multi-agent computer science literature, in **Chapter 4** I develop a new class of multi-agent games which aim to incorporate some of the complexity and diversity of real life with the formal precision of traditional economic games. These games can be played intuitively by people.

Empirically, I find that anonymously matched people robustly reciprocate even when the game changes after each interaction. People can infer whether others intend to cooperate or compete after observing just a single ambiguous movement and quickly reciprocate the inferred intention. In new environments, people generalize abstract intentions like cooperation and competition by executing a novel set of low-level movements needed to realize those goals. Finally, many dyads develop roles and norms after a few interactions that increase the efficiency of cooperation by coordinating their actions. These novel empirical findings both demonstrate the power of human social cognition and are the challenge for computational models to explain and replicate.

To understand and predict human behavior in these games I develop a novel model that treats cooperation and competition as probabilistic planning programs. To realize cooperation algorithmically, I formalize, for the first time, an influential psychological account of collaboration known as "joint intentionality." In our model, each agent simulates a mental model of the group (oneself included) from an impartial view. From this view the group itself is treated as a single agent with joint control of each individual and with the aim of optimizing a shared goal. An agent then plays its role in this joint plan leading to the emergence of roles. Competition is realized by iterating a best response to the inferred intention of the other player.

These models of abstract cooperation and competition serve a dual role: they are abstract models of cooperative and competitive action and also the likelihood in a hierarchical Bayesian model that infers whether or not other agents are cooperating. This inference realizes a sophisticated form of theory of mind. With these pieces of cognitive machinery in place, reciprocity is realized by mirroring the inferred intentions of the other players. This model explains the key empirical findings and is a first step towards understanding the cognitive microstructure of cooperation in terms of rational inference and multi-agent planning.

In **Chapter 5** I develop a novel scheme for probabilistic inference over an infinite space of possible strategies. Inferring underlying cooperative and competitive strategies from human behavior in repeated games is important for accurately characterizing human behavior and understanding how people reason strategically. Finite automata, a bounded model of computation, have been extensively used to compactly represent strategies for these games and are a standard tool in game theoretic analyses. However, inference over these strategies in repeated games is challenging since the number of possible strategies grows exponentially with the number of repetitions yet behavioral data is often sparse and noisy. As a result, previous approaches start by specifying an finite hypothesis space of automata which does not allow for flexibility. This limitation hinders the discovery of novel strategies which may be used by humans but are not anticipated a priori by current theory.

I present a new probabilistic model for strategy inference in repeated games by exploiting non-parametric Bayesian modeling. With simulated data, I show the model is effective at inferring the true strategy rapidly and from limited data which leads to accurate predictions of future behavior. When applied to experimental data of human behavior in a repeated prisoners dilemma, I uncover new strategies of varying complexity and diversity.

In **Chapter 6** I study how humans allocate the spoils of a cooperative endeavor expanding the scope of cooperation to cases where benefits are unequally distributed. Lasting cooperation depends on allocating those benefits fairly according to normative principles. Empirically I show that in addition to preferences over outcomes such as the efficiency and equitability of a distribution, we are also sensitive to the attributions others might make about us as a result of our distribution decisions. We care about our reputations and whether we will be seen as trustworthy and impartial partners in the future.

Preferences of this type require reasoning about and anticipating the beliefs others will form as a result of one's action. To explain these results I develop a model which integrates theory of mind into a utility calculus. By turning the cognitive capacity to infer latent desires and beliefs from behavior towards oneself, agents anticipate the judgments others will make about them and incorporate those anticipated judgments as a weighted component of an agent's utility function. Across many scenarios tested with behavioral experiments my model quantitatively explains both how people make hypothetical resource allocation decisions and the degree to which they judge that others who made decisions in the same contexts as impartial. These empirical results understood through our model, shed light on the ways in which our cooperative behavior is shaped by the desire to signal prosocial orientations.

Finally, in **Chapter 7** I study the computational structure of moral judgment. One puzzle of moral judgment is that while moral theories are often described in terms of absolute rules (e.g., the greatest amount of good for the greatest number, or the doctrine of double effect), our moral judgments are graded. Since moral judgments are particularly sensitive to the agent's mental states, uncertainty in these inferred mental states might partially underlie these graded responses. I develop a novel computational representation for reasoning about other people's intentions based on counterfactual contrasts over influence diagrams. This model captures the future-oriented aspect of intentional plans and distinguishes between intended outcomes and unintended side effects a key feature needed for moral judgment.

I give a probabilistic account of moral permissibility which produces graded judgments by integrating uncertainty about inferred intentions (deontology) with welfare maximization (utilitarian). By grounding moral permissibility in an intuitive theory of planning, I quantitatively predict the fine-grained structure of both intention and moral permissibility judgments in classic and novel moral dilemmas. In an era of autonomous vehicles and, more generally, autonomous AI agents that interact with or on behalf of people, the issue has now become relevant to AI as well. My models point towards ways to imbue AI agents with some means for evaluating the moral responsibility of their own actions as well as the actions of others.

I conclude in **Chapter 8** and discuss the implications of this work for understanding the social mind.

- Chapter 2 is based on Kleiman-Weiner, M., Vientós, A., Rand, D.G., & Tenenbaum, J. B. (submitted). The Evolution of Cooperation in Cognitively Flexible Agents.

- Chapter 3 is based on: Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition.*

- Chapter 4 is based on: Kleiman-Weiner, M., Ho, M., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

- Chapter 5 is based on: Kleiman-Weiner, M., Tenenbaum, J. B., & Zhou, P. (in press). Non-parametric Bayesian inference of strategies in infinitely repeated games. *Econometrics Journal.*

- Chapter 6 is based on: Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing Social Preferences From Anticipated Judgments: When Impartial Inequity is Fair and Why? *Proceedings of the 39th Annual Conference of the Cognitive Science Society.*

- Chapter 7 is based on: Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Chapter 2

The Evolution of Cooperation in Cognitively Flexible Agents

Despite great advances in our understanding of the evolution of cooperation through reciprocity (Axelrod, 1985; M. A. Nowak, 2006; Rand & Nowak, 2013), a central aspect of human cooperation has not been explained: our flexibility. In real life (unlike a repeated prisoner's dilemma), each social interaction is unique and complex; we never play the same game twice. Our social cognition allows us to cooperate (or withhold cooperation) in an infinitude of environments with known and unknown partners in novel situations where we must trade-off our own interests with the interests of others.

This intelligence also lets us act in situations where we have nothing directly at stake, such as allocating finite or scarce resources among several others, or even judge hypothetical moral dilemma, such as whether to save the life of many at the cost of sacrificing one. Distinctively human cooperation also requires abstraction: we learn and plan with abstract moral principles that determine how the benefits of cooperation should be distributed and how those who fail to cooperate should be treated (Kleiman-Weiner, Saxe, & Tenenbaum, 2017). These theories enable us to generalize from sparse noisy behavior in specific situations to abstract knowledge of who we should cooperate with and how. Even young children rapidly evaluate and abstractly reciprocate based on the actions of others (Warneken & Tomasello, 2006; Hamlin, Wynn, & Bloom, 2007; Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Choi & Luo, 2015).

Computer scientists and economists have developed formal quantitative frameworks to try to understand cooperation, game theory being a prominent example (C. Camerer, 2003; Rand & Nowak, 2013). However, these frameworks do not capture some of the most interesting aspects of human cooperation. Compared to behavioral automata (such as Tit-For-Tat) that are hand-designed for cooperation in a single task, or reinforcement learning algorithms that require long periods of trial-and-error learning, people cooperate much more flexibly with much less experience.

For example, in prisoner's dilemma two players each choose one of two actions, Cooperate (C) or Defect (D) with consequences such that if both players make the same choice, they are better off cooperating, but either player can individually gain most in a given interaction by defecting when the other cooperates. If players play each other many times repeatedly then a stable strategy under evolutionary dynamics is Tit-for-Tat (Axelrod, 1985; M. A. Nowak, 2006), in which a player always begins any new interaction by cooperating, and on subsequent rounds play whatever the other player played last: cooperating if the other player last cooperated, otherwise defecting. This starts to look like reciprocity: trusting strangers and being altruistic towards them, inviting cooperation but exposing themselves to exploitation; punishing any betrayal of that trust in an eye-for-an-eye fashion; but being forgiving if the other player returns to cooperative behavior.

While our goal here is similar to these evolutionary models which we feel have much to offer, our approach to learning – both *what* is learned, and *how* it can be learned – is importantly different, and begins where these behavioral models typically leave off. The sense in which agents in typical evolutionary simulations actually express any principles of genuine cooperation, is very limited and strictly dependent on our own human intuitions for interpreting strategies such as Tit-for-Tat. They look cooperative chiefly because that is how we humans most naturally explain their behavior. What agents in those simulations have actually learned is merely to choose a particular row or column in a single matrix representing a particular game's payoff structure.

If confronted with an even slightly different game representing a slightly different decision, nothing that agents in a typical evolutionary simulation have learned generalizes at all. For example, agents who cooperate in prisoner's dilemma – that is, to choose the *C* row or column in a 2×2 (or $[C,D] \times [C,D]$) matrix – haven't learned to be altruistic in dictator games or to be trusting in public goods games, even though these are all very similar (Peysakhovich & Rand, 2015).

In this work, we address these shortcomings head-on and develop a new cognitive model of agency that captures the flexibility of distinctively human cooperation. Our model is based on an abstract and reciprocal utility calculus where agents value the welfare of others who they believe share their "values" i.e., behave under the same reciprocal utility calculus. Since "values" cannot be observed directly they must be inferred from observations of behavior. This is particularly complicated in these social games because attributing intentions to an agent's actions is highly overdetermined. When we observe another agent defect it might have the intention to retaliate for a previous defection, on accident, or because that agent is simply selfish. To handle this uncertainty we develop a novel variant of Bayesian theory-of-mind that allows agents to reason directly about "joint-belief," or what groups of agents know together.

To test these models we develop a new challenge task for evolutionary games theory called *dynamic dilemmas* where each social interaction is uniquely generated from a probabilistic distribution. Within this space of games are familiar stage games such as the prisoner's dilemma, altruistic giving games where players can give up some of their own welfare to help another person, allocation games where players can show favoritism in choosing who should receive an indivisible resource, and even moral dilemmas where players bear no personal costs themselves but have to decide outcomes for groups of others. Since each interaction is sampled probabilistically, no two interactions are ever exactly alike.

With these new tools we investigate the evolutionary dynamics of our agents in the *dynamic dilemma* and show our models flexibly reciprocate in this new task outcompeting selfish agents. We show that since this model learns abstractly it can effectively cooperate in situations where it learns directly from its own experience (direct reciprocity) as well as situations where it must learn only from observation (indirect reciprocity). Finally, we compare our model to existing behavior models in the repeated prisoner's dilemma and show that our cognitive agents out-compete the leading finite state automata including Tit-for-Tat and Win-Stay-Lose-Shift. We conclude by discussing the implications of this work for understanding the origins of uniquely human social cognition and its value for enabling cooperation.

2.1 Formal Model

We first describe a novel mathematical framework that combines computational cognitive modeling with the tools of evolutionary game theory. Unlike previous work, this framework allows us to study flexible cooperation as well as the cognitive computations and representations which underlie it. We borrow the notation of multi-agent utility based decision making.

In general, at time t an agent i is faced with a decision to choose an action a from a set of actions $a \in \mathcal{A}_t$. Each action specifies a distribution of welfare ($\mathbf{R}$) to the other agents where $R_i(a)$ is the distribution of welfare to agent i caused by action a. For each action, there is a probability ε that a random action is taken instead of the agent's chosen action ("trembling hand"). Finally, each action is observed by a subset of other agents (O_t). The set of alternative actions, probability of randomness, and set of observers are stored in the state $s_t = (\mathcal{A}_t, \varepsilon, O_t)$. The observation of (s_t, a_t) is assumed to be common knowledge for all observers in O_t.

2.1.1 Recursive Reciprocal Utility

We start with the basic assumption that intelligent agents have a utility function which they aim to maximize. This simple engineering principle of intelligence is usually absent in the behavioral models used in evolutionary game theory. The simplest version of this idea is an agent that aims to maximize only the reward it receives:

$$U_i^{\texttt{selfish}}(a) = R_i(a) \tag{2.1}$$

where i only values her own welfare. This utility function is identical to the utility functions commonly used in single agent reinforcement learning tasks. Since an agent with this utility function doesn't directly value the welfare of anyone else, we call agents with this utility function `selfish`.

In order to explain altruistic behavior, other utility functions have been studied where an agent values the welfare of not only herself but also the welfare of others (Kiley Hamlin et al., 2013; Kleiman-Weiner, Saxe, & Tenenbaum, 2017). One simple way to realize this idea computationally is by agents recursively valuing the welfare of other agents:

$$U_i^{\texttt{altruistic}}(a) = R_i(a) + \sum_{j \neq i} R_j(a) \tag{2.2}$$

Since an agent with this utility function will act towards the welfare of all we call agents with this utility function `altruistic`.

When compared to the simple agents traditionally studied in the evolutionary game theory, agents with a `selfish` or `altruistic` utility function correspond to a more sophisticated version of the Always-Defect (AllD) and Always-Cooperate strategies respectively. In a repeated prisoner's dilemma, our agents act identically to their corresponding behaviorist strategies but generalize their behavior to any situation or context that can be evaluated using a utility function.

While an `altruistic` agent will act cooperatively, it is not sufficient for the evolution of cooperation since it cooperates *unconditionally* and thus can be exploited by `selfish` agents. In order to understand how cooperation might stably evolve we seek to build agents that cooperate *conditionally* i.e., reciprocally. The Tit-for-Tat (TFT) strategy that emerged as the winner of Alexrod's 1981 contest is a celebrated example of a simple strategy that reciprocates based on its partner's *behavior* (Axelrod, 1985). In an evolutionary tournament, TFT avoids exploitation by AllD while still maintaining cooperation with itself and the other cooperative strategies.

In contrast to previous work which encodes reciprocal behavior in rules, we develop an agent we call `reciprocal` which has a simple recursive utility function that gives rise to a reciprocal preference:

$$U_i^{\texttt{reciprocal}}(a) = R_i(a) + \sum_{j \neq i} R_j(a) \mathbb{1}(U_j = U_i)$$

where $\mathbb{1}(U_j = U_i)$ is 1 if j has the same utility function as i and 0 otherwise. Thus if j and i have the same utility function, i will weight the welfare of j equal to its own in its decision making process and otherwise will ignore the welfare of j. Unlike previous approaches such as TFT and its variants which reciprocate cooperation behaviorally, the `reciprocal` agent reciprocates cognitively: it values the welfare of those who share its system of reciprocal valuation.

However, unlike behavior, the utility functions of other agents cannot be observed directly and thus the agent cannot calculate $\mathbb{1}(U_j = U_i)$. Instead utility functions (and the values they encode) must be inferred through behavior. Replacing $\mathbb{1}(U_j = U_i)$ in the `reciprocal` agent with its expectation:

$$U_i^{\texttt{reciprocal}}(a,s,B) = R_i(a) + \sum_{j \neq i} R_j(a) B(U_j = U_i) \tag{2.3}$$

where $B(U_j = U_i)$ are i's beliefs that i and j share the same utility function. If i believes that j is of the same type as i then it will weight the welfare of j highly in its decision making process. Likewise if i infers j is not likely to share its utility function, it will not value j's welfare highly or at all.

These beliefs can be computed rationally based on i's past observations H^i) using Bayesian inference. Since the utility function of the `reciprocal` agent depends on the interactions it has observed, it will act differently in the same situation depending on the inferences it has made. This approach is intentional by design i.e., it correctly reasons about accidental actions and false beliefs and is character-driven: the `reciprocal` agent doesn't evaluate an action in and of itself, instead it evaluates what that action reveals about the actor's underlying utility function (D. A. Pizarro & Tannenbaum, 2011; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015).

Decision-making is then defined probabilistically under the Luce-choice decision rule subject to the agent's utility function U and beliefs B. This soft-maximization reflects utility maximization when there is uncertainty about the exact utility value:

$$P_B(a|s,U) \propto e^{\beta U(a,s,B)} \tag{2.4}$$

When $\beta \to 0$ the decision maker chooses randomly and when $\beta \to \infty$ the decision maker will always choose an action with the highest utility. Beliefs (B) are ignored for the `selfish` and `altruistic` utility functions since their utility functions do not depend on their belief.

2.1.2 Inferring Recursive Social Knowledge

We now show how agents can rationally learn the types of other agents based on observations of behavior using Bayesian inference. Let $H_t^i = \{(s_1, a_1), \dots (s_t, a_t)\}$ be the sequence of observations observed by i up until time t. Beliefs (B) about the mental states (desires and knowledge) of others can can be written as a Bayesian inference (Liddle & Nettle, 2006; C. L. Baker et al., 2009; Ullman et al., 2009; C. L. Baker et al., 2017; Kleiman-Weiner et al., 2015; Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Kleiman-Weiner, Saxe, & Tenenbaum, 2017):

$$P(U_j|H_t^i) \propto P(H_t^i|U_j)P(U_j) \tag{2.5}$$

the posterior beliefs on the utility function of j, U_j where $P(U_j)$ is the prior over a space of possible utility functions. How to compute the likelihood $P(H_t^i|U_j)$? First, the posterior at time t can be computed directly from our posterior at time $t-1$:

$$P(U_j|H_t^i) \propto P_{B_{t-1}}(a_t|s_t, U_j)P(U_j|H_{t-1}^i) \tag{2.6}$$

$P_{B_{t-1}}(a_t|s_t, U_j)$ denotes the likelihood of action a_t being taken by an agent of type U_j given their beliefs B at time $t-1$. To compute this likelihood, i must have a model of what j's beliefs were at time $t-1$, when a_t took place. These models must be recursive. For instance, when i observes j defect on k, he must assess j's action in terms of what j believes about k. If j and i have seen k defect on many others, then j's defection might be seen as reciprocity rather than selfishness. Such beliefs can be computed if i has models of every other agent, which in turn have their own models of every other agent, and so on (Yoshida, Dolan, & Friston, 2008; Ullman et al., 2009; Kleiman-Weiner et al., 2016). However, this approach scales poorly as the number of models (and hence model updates)

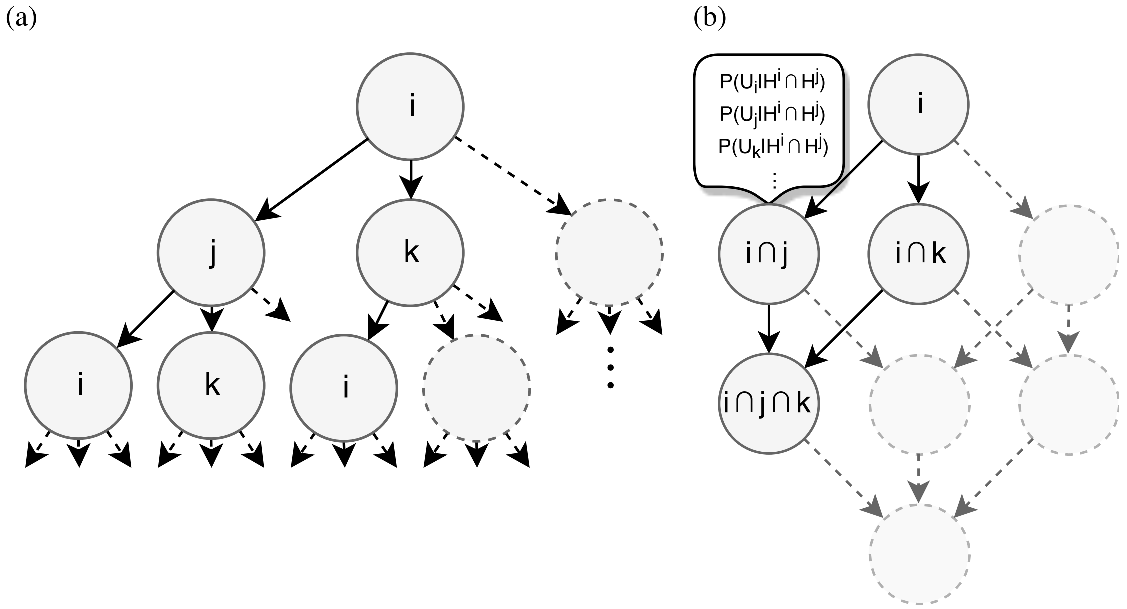


Figure 2-1: Recursive representations of social knowledge. (a) Agent *i* has a model of each agent which recursively has a model of each other agent including *i*. (b) Agent *i* reasons directly about the joint belief of each group of agents. Pop-out shows an example of the joint-beliefs for each subset.

grows exponentially with the number of interactions observed by each agent which makes it intractable to compute in a repeated game (Figure 2-1a). When approximated with a finite depth, beliefs quickly diverge and are unstable.

We develop a new approach for modeling higher-order theory-of-mind that is both efficient and does not require approximations. The key insight is that *i*'s model of *j*'s beliefs only depends on the events jointly observed by *i* and *j*. If an event was observed by *i* and not *j*, or was observed by *j* and not *i* it can't be part of *i*'s model of *j*'s beliefs because in the first case it would be irrational and in the second case *i* would not be aware of it. As a result, *i*'s model of *j*'s beliefs are identical to the beliefs an external agent in a "view from nowhere" would form had they observed what was jointly observed by *i* and *j* (Nagel, 1989). We call this knowledge a *joint-belief*; it is *i*'s model of what he and *j* know together (Fagin, Halpern, Moses, & Vardi, 2004; Gmytrasiewicz & Durfee, 2000). Essentially when we ask what does a subset of agents know about each other, the joint-belief partitions the observations to only include the set of observations that were jointly observed by that subset. Note: we are not suggesting that there is anything actually external to the minds of

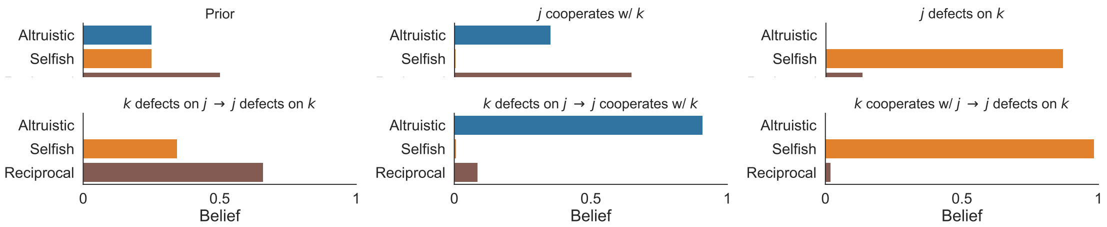


Figure 2-2: Belief updates using recursive theory-of-mind. Each graph shows i's beliefs about j in response to the scenario written above. Note that j's same action leads to different beliefs depending on what both i and j have observed about k. Here, cooperation corresponds to paying 1 welfare to give another person 3 welfare and defection corresponds to doing nothing.

individual agents, but rather that each agent itself can reason about the beliefs they hold jointly with others.

Formally, joint-belief updates about j's utility function are written recursively:

$$P(U_j|H_t^i) \propto P_{P(U|H_{t-1}^i \cap H_{t-1}^j)}(a_t|s_t,U_j)P(U_j|H_{t-1}^i) \tag{2.7}$$

where $P_{P(U|H_{t-1}^i \cap H_{t-1}^j)}(a_t|s_t,U_j)$ is the likelihood of action shown in equation (2.4) with $P(U|H_{t-1}^i \cap H_{t-1}^j)$ as the belief state. $H_{t-1}^i \cap H_{t-1}^j$ are the joint observations of agents i and j that were made before a_t. Because $H_{t-1}^i \cap H_{t-1}^j = H_{t-1}^j \cap H_{t-1}^i$ we no longer have to represent a growing hierarchy of models (Figure 2-1b). By reasoning about joint-belief, each agent does not need to represent every permutation of i believes j believes k and so on. Instead each agent represent what i, j, and k jointly believe which can be computed efficiently using dynamic programming. In general, the joint-belief representation greatly simplifies reasoning about knowledge when a large group of agents observe new information. Instead of updating what each agent knows individually (and what they believe others know etc.), joint-belief allows agents to directly represent the joint knowledge of the group, and update that joint knowledge all at once. Figure 2-2 shows some example interactions between i, j, and k and how i's joint-beliefs update over in response to observations. Without a recursive theory of mind, agents would not be able to differentiate a defection done for retaliatory reasons from a defection done for selfish reasons. The only thing differentiating these situations are joint-beliefs.

2.2 Evolutionary Analysis

In the spirit of rational analysis, our goal is to understand if and when the abstract reciprocity we have proposed and the inference that supports it is adaptive (Marr, 1982). While basic forms of perception (e.g., object detection) can be shown as adaptive with respect to a non-cognitive environment, social cognition must be evaluated against an environment of other cognitive agents who themselves may be learning and adapting their behavior. For instance, if measured in isolation, a group of `altruistic` agents generate more surplus welfare than a group of `reciprocal` agents. However, this analysis would not take into account the fact that the introduction of a single `selfish` agent would exploit the group of `altruistic` agents but would be detected and rebuffed by the group of `reciprocal` agents. Therefore we need to also compare the stability of our agents in the presence of others in addition to how effective they are at generating surplus welfare. Evolutionary analysis gives us this property and is also suggestive of a mechanism for how the types of agents might change over time (either through cultural learning or biological evolution) (M. A. Nowak, 2006; Sigmund, 2010; Rand & Nowak, 2013).

Simulation Details Simulation of a selection process, like a Markov process such as MCMC (Suchow, Bourgin, & Griffiths, 2017), amounts to finding the steady state frequencies of each agent type after many generations. We follow the best practices described in Sigmund (2010). In short, N agents interacted, one agent was chosen at random and selected another agent with probability proportional to its cumulative welfare ($\propto \exp(s\dot{R})$) where s is the strength of selection. There is also a small rate of mutation where with probability δ the agent just picks a new utility function from a space of possible functions. This is called the Moran Process and for most of the analyses here, we show steady state frequencies of this process in the low mutation limit ($\delta \to 0$). See Sigmund (2010) for the mathematical details of calculating these frequencies.

The agents in our simulations used parameters $\beta = 5$, a prior of $P(U = U^{\texttt{reciprocal}}) = 0.5$ with the remaining 0.5 spread uniformly across all other agent types. The types of agents considered for theory-of-mind inference was always equal to the actual agents in that simulation. These priors can be learned as well which do not have room to describe here.

The Dynamic Dilemma is a generator for a series of interactions where agents never have the same interaction twice. Each interaction is generated by matching up a random subset of at least two players and randomly assigning one player as the decision maker for that scenario. Each pair of individuals might be sampled again with probability γ so the expected number of interactions between each pair of agents is $1+\frac{1}{1-\gamma}$.

The decision maker is faced with a set of choices $\mathcal{A}$ which always includes the option of doing nothing. Doing nothing corresponds to giving zero welfare to oneself and all the other players one is interacting with. For each other paired partner the decision maker can pay a cost (c) which reduces its welfare to deliver a benefit (b) which increases the welfare of another player. The c are sampled from a Poisson distribution so that sometimes there is no cost. The b are generated by sampling two random exponential variables (e_0 and e_1) and setting $b = e_0 c + e_1$ which enforces $b > c$. To determine the number of actions in a game, we sample c and b a (1+$N_{\mathcal{A}}$) number of times which is itself sampled from a Poisson distribution. For each $N_{\mathcal{A}}$ we sample a new c and b and for each of these samples create a set of choices where each choice corresponds to one of the non-deciding players receiving b while the decision maker pays c.

The *dynamic dilemma* environment has stochasticity. Each interaction has a probability of a "trembling hand" where with sampled probability ε, a random action is taken instead of the action chosen. Finally, each interaction is observed by all agents with probability ω and is otherwise only observed by those matched for that specific interaction. For the simulations shown here in we matched agents into groups of 2 or 3 with equal probability. For the other parameters we set $\lambda_c = 1$, $\lambda_{N_{\mathcal{A}}} = 1$, $\lambda_{e_0} = \lambda_{e_1} = 5$ and $\varepsilon = 0.01$. Figure 2-3 shows a visual depiction of the game engine which produces these dilemma.

No two situations generated in the *dynamic dilemma* are exactly alike yet well known games correspond to particular regions of the parameter space. For instance, when c and b are constant, the number of players are limited to two players, $N_{\mathcal{A}} = 1$, and $\omega = 0$ we can recover a sequential or simultaneous prisoner's dilemma depending on the temporal structure of observation. If observation happens only after both players have acted then

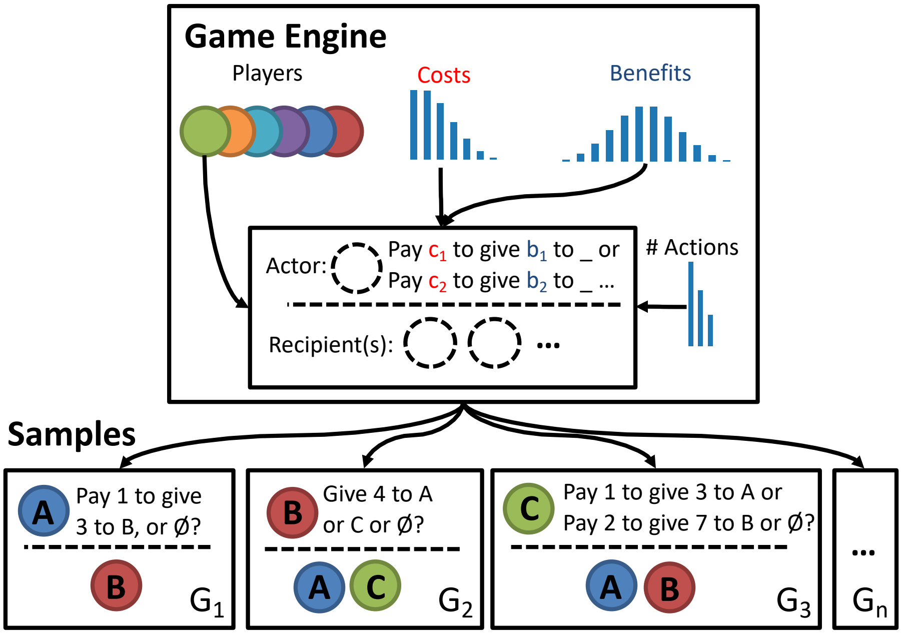


Figure 2-3: Visual depiction of the generative "game engine" which generates unique dilemma. The top shows the ingredients: a set of players, distributions over costs, benefits, and number of actions and a template which describes how these elements are combined into a choice dilemma. Below show example samples which were generated from the game engine.

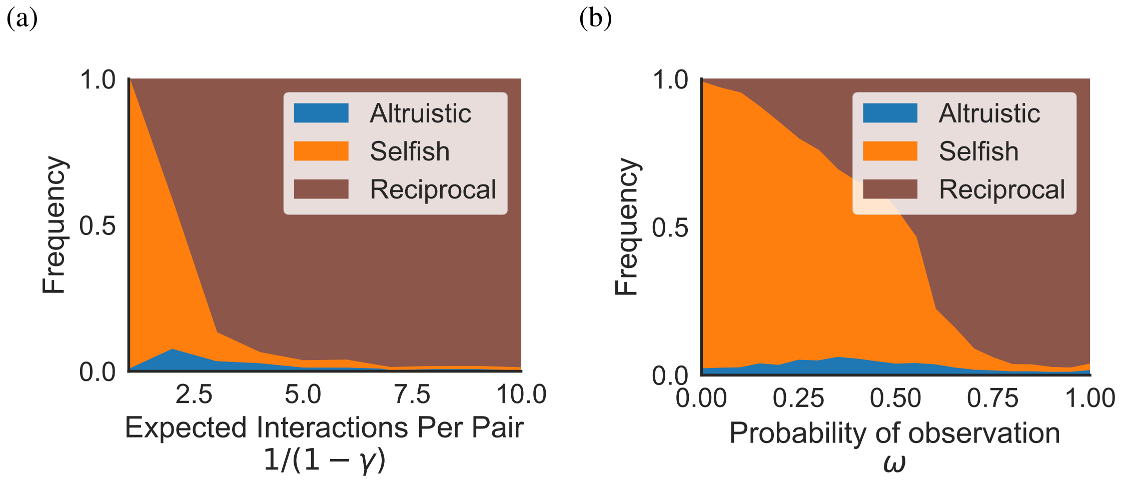


Figure 2-4: Evolutionary analysis of the `dynamic dilemma` showing the steady state frequencies of the three agent types as a function of the expected number of repeated interactions (a) and the probability that everyone will observe a given interaction when agents only interact once (b). Simulations were run with $N = 10$ agents and a selection strength of $s = 0.5$

the game is a traditional prisoner's dilemma, if observation happens after only one of the players decides it becomes a sequential prisoner's dilemma. When the number of players is fixed to three and $c = 0$, the agents play an allocation game, choosing how to distribute an indivisible resource.

2.3 Results

Evolution of Cognitively Flexible Cooperation Using simulations we study the evolution of the `reciprocal` agent compared against the `selfish` and `altruistic` agents in the *dynamic dilemma*. Figure 2-4a shows the steady state frequency of the three agent types. When the expected number of repeated interactions γ is low, the `selfish` agent dominates. However, as the expected number of repeated interactions grows, the `reciprocal` agent becomes the dominant agent in the population.

Next, we investigated how the probability of public observation impacts the equilibrium frequencies of the agents even when each pair of agents only interact once. Figure 2-4b shows that when the probability of public observation was low $\omega < 0.5$ the `selfish` agent dominates since the `reciprocal` agent does not receive enough information to correctly infer types. For higher ω, the `reciprocal` agent is able to learn who to cooperate with before interacting.

The `reciprocal` agent unifies both direct and indirect forms of reciprocity in the challenging *dynamic dilemma*. When posed as a problem of inference, the behavioral observations obtained through direct interactions and those from observing the interactions of others lead to the same types of belief updates. In reality this will not always be true. For instance, in an indirectly observed interaction, the space of possible alternatives might not be as obvious as when one is directly interacting so inference might be less accurate. We leave this as a challenge for future work.

Evolution Favors Cognition over Behavioral Rules Since the `reciprocal` agent can play any game, it can also compete in a repeated prisoner's dilemma tournament (Axelrod, 1985). In Figure 2-5 we ran the `reciprocal` agent against the dominant finite state automata strategies from the literature. In addition to TFT, AllD and AllC (discussed above), we added: generous-TFT (GTFT) a forgiving variant that has a probability of spontaneously returning to cooperation after a defection, Win-Stay-Lose-Shift (WSLS, also called Pavlov) which treats mutual cooperation and defecting against cooperation as "wins" and mutual defection and cooperation against defection as a "loss" (M. A. Nowak, 2006).

Our first set of simulations replicate the main results from the literature. When there is no noise in the environment ($\epsilon = 0$), and there is sufficient repetition TFT (red) is most favored by evolution (Figure 2-5a). However, no single strategy is dominant. Since TFT also cooperates with AllC (blue), AllC is always present at a low frequency. The presence of these unconditional cooperators is an avenue for AllD (orange) to return and so there is a heterogeneous mixture of strategies at in the steady state. In contrast, when the `reciprocal` agent (brown) is added to the tournament (Figure 2-5b), it out-competes all of the automata strategies and prevents a heterogeneous mixture of agents by not cooperating with agents it has inferred are unconditional cooperators.

We next investigated a stochastic environment where with probability ϵ each agent takes a random action (M. A. Nowak, 2006). The dominant strategy from the first tournament, TFT, is not robust to noise since even a single perturbed action can lead TFT to alternate

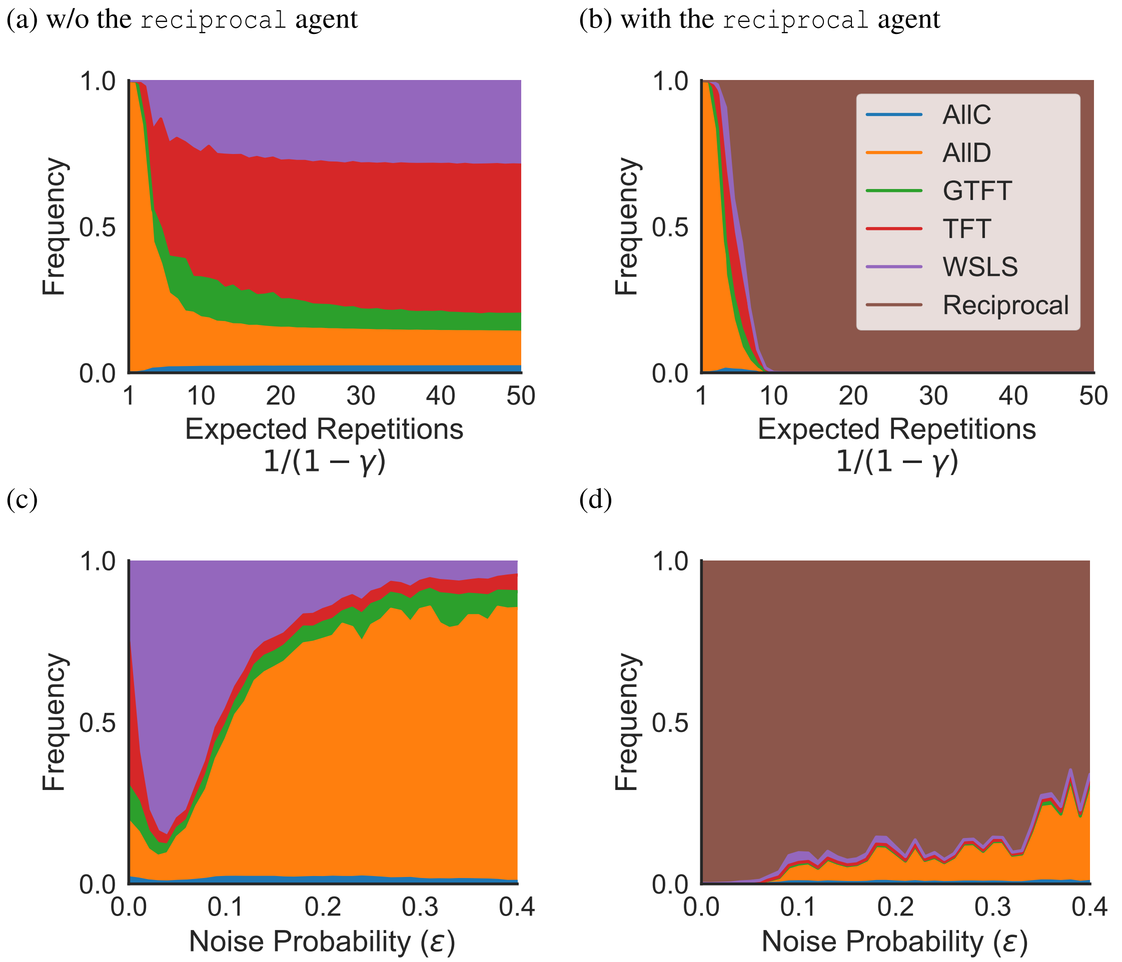


Figure 2-5: Evolutionary analysis of an Axelrod style repeated prisoner's dilemma (RPD) tournament for different parameter settings: expected number of rounds with no noise (top) and the probability of noisy action with 10 expected rounds ($\gamma = 0.9$) (bottom). We compare tournaments that only include the automata strategies (left) with those that also include the `reciprocal` agent (right). The resulting steady state frequency of each agent type is plotted as stacked bars since the total frequency must sum to one. For all the tournaments each stage game had $\frac{b}{c} = 3$ and selection was run with $N = 100$ agents and selection strength $s = 0.5$.

between cooperation and defection instead of stably cooperating with itself. Consistent with the literature (M. A. Nowak, 2006), in these stochastic environments WSLS (purple) dominates since it has an error correction mechanism (Figure 2-5c). However, it is only robust for relatively low amounts of noise. As the probability of noise increases, AllD takes over again. Again, when we add the `reciprocal` agent to the tournament it robustly evolves even at high noise levels (Figure 2-5d).

Unlike automata strategies where different strategies perform best in certain environments (TFT when there is no noise and WSLS when there is some noise), the `reciprocal` agent performed well across a range of environment. Surprisingly, the `reciprocal` agent outperformed the automata strategies even though these automata were hand-designed for cooperation in this specific context and the `reciprocal` agent was designed for flexible cooperation in any context.

2.4 Conclusion

Our work is a first formal investigation of how computational principles of cognition such as abstraction and inference enable the scale and scope of human cooperation. These simulations give insights into the origins of distinctively human cooperation and how they depend on distantly human forms of social cognition (Tomasello, 2014). For instance how special representations for reasoning about the knowledge of others such as “joint-belief” can be formed through a recursive theory-of-mind. Deeper still, these studies provide a start for understanding the evolutionary origins of social cognition. Using this new framework we can precisely quantify when social cognition delivers benefits above and beyond less sophisticated and less flexible agents (Singh, Lewis, Barto, & Sorg, 2010). This is a step towards rigorously understanding the ultimate origins of our social intelligence and how it both supports and is supported by distinctively human cooperation.

Chapter 3

Learning a Commonsense Moral Theory

> *Common sense suggests that each of us should live his own life (autonomy), give special consideration to certain others (obligation), have some significant concern for the general good (neutral values), and treat the people he deals with decently (deontology). It also suggests that these aims may produce serious inner conflict. Common sense doesn't have the last word in ethics or anywhere else, but it has, as J. L. Austin said about ordinary language, the first word: it should be examined before it is discarded.*
>
> – Thomas Nagel, The View From Nowhere

Basic to any commonsense notion of human morality is a system of values for trading off the interests and welfare of different people. The complexities of social living confront us with the need to make these trade-offs every day: between our own interests and those of others, between our friends, family or group members versus the larger society, people we know who have been good to us or good to others, and people we have never met before or never will meet. Morality demands some consideration for the welfare of people we dislike, and even in some cases for our sworn enemies. Complex moral concepts such as altruism, fairness, loyalty, justice, virtue and obligation have their roots in these trade-offs, and children are sensitive to them in some form from an early age. Our goal in this paper is to provide a computational framework for understanding how people might learn to make these trade-offs in their decisions and judgments, and the implications of possible learning mechanisms for the dynamics of how a society's collective morality might change over time.

Although some aspects of morality may be innate, and all learning depends in some form on innate structures and mechanisms, there must be a substantial role for learning from experience in how human beings come to see trade-offs among agents' potentially conflicting interests (Mikhail, 2007, 2011). Societies in different places and eras have differed significantly in how they judge these trade-offs should be made (Henrich et al., 2001; P. Blake et al., 2015; House et al., 2013). For example, while some societies view preferential treatment of kin as a kind of corruption (nepotism), others view it as a moral obligation (what kind of monster hires a stranger instead of his own brother?). Similarly, some cultures emphasize equal obligations to all human beings, while others focus on special obligations to one's own group e.g. nation, ethnic group, etc. Even within societies, different groups, different families, and different individuals may have different standards (Graham, Haidt, & Nosek, 2009). Such large differences both between and within cultures pose a key learning challenge: how to infer and acquire appropriate values, for moral trade-offs of this kind. How do we learn what we owe to each other?

Children cannot simply learn case by case from experience how to trade off the interests of specific sets of agents in specific situations. Our moral sense must invoke abstract principles for judging trade-offs among the interests of individuals we have not previously interacted with or who have not interacted with each other. These principles must be general enough to apply to situations that neither we nor anyone we know has experienced. They may also be weighted, such that some principles loom larger or take precedence over others. We will refer to a weighted set of principles for how to value others as a "moral theory," although we recognize this is just one aspect of people's intuitive theories in the moral domain.

The primary data that young children observe are rarely explicit instructions about these abstract principles or their weights (J. C. Wright & Bartsch, 2008). More often children observe a combination of reward and punishment tied to the moral status of their own actions, and examples of adults making analogous decisions and judgments about what they (the adults) consider morally appropriate trade-offs. The decisions and judgments children observe typically reflect adults' own moral theories only indirectly and noisily. How do we generalize from sparse, noisy, underdetermined observations of specific instances of moral behavior and judgment to abstract theories of how to value other agents that we can then apply everywhere?

Our main contribution in this paper is to posit and formalize a minimal set of cognitive capacities that people might use to solve this learning problem. Our proposal has three components:

- **An abstract and recursive utility calculus.** Moral theories (for the purposes of trading off different agents' interests) can be formalized as values or weights that an agent attaches to a set of abstract principles for how to factor any other agents' utility functions into their own utility-based decision-making and judgment.
- **Hierarchical Bayesian inference.** Learners can rapidly and reliably infer the weights that other agents attach to these principles from observing their behavior through mechanisms of hierarchical Bayesian inference; enabling moral learning at the level of values on abstract moral principles rather than behavioral imitation.
- **Learning by value alignment.** Learners set their own values guided by meta-values, or principles for what kinds of values they value holding. These meta-values can seek to align learners' moral theories externally with those of others ("We value the values of those we value"), as well as internally, to be consistent with their own attachments and feelings.

Although our focus is on the problems of moral learning and learnability, we will also explore the implications of our learning framework for the dynamics of how moral systems might change within and across generations in a society. Here the challenges are to explain how the same mechanisms that allow for the robust and stable acquisition of a moral theory can under the right circumstances support change into a rather different theory of how others interests are to be valued. Sometimes change can proceed very quickly within the span of one or a few generations; sometimes it is much slower. Often change appears to be progressive in a consistent direction towards more universal, less parochial systems

– an “expanding circle” of others whose interests are to be taken into account, in addition to our own and those of the people closest to us (Singer, 1981; Pinker, 2011). What determines when moral change will proceed quickly or slowly? What factors contribute to an expanding circle, and when is that dynamic stable? These questions are much bigger than any answers we can give here, but we will illustrate a few ways in which our learning framework might begin to address them.

The remainder of this introduction presents in more detail our motivation for this framework and the phenomena we seek to explain. The body of the paper then presents one specific way of instantiating these ideas in a mathematical model, and explores its properties through simulation. As first attempts, the models we describe here, though oversimplified in some respects, still capture some interesting features of the problems of moral learning, and potential solutions. We hope these features will be sufficient to point the way forward for future work. We conclude by discussing what is left out of our framework, and ways it could be enriched or extended going forward.

The first key component of our model is the expression of moral values in terms of utility functions, and specifically recursively defined utilities that let one agent take others’ utilities as direct contributors to their own utility function. By grounding moral principles in these recursive utilities, we have gained a straightforward method for capturing aspects of moral decision-making in which agents take into account the effects of their actions on the well-being of others, in addition to (or indeed as a fundamental contributor to) their own well-being. The specifics of this welfare are relatively abstract. It could refer to pleasure and harm, but could also include other outcomes with intrinsic value such as “base goods” e.g., achievement and knowledge (Hurka, 2003) or “primary goods” e.g., liberties, opportunities, income (Rawls, 1971; Scanlon, 1975; Sen & Hawthorn, 1988) or even purity and other “moral foundations” (Haidt, 2007). This proposal thus formalizes an intuitive idea of morality as the obligation to treat others as they would wish to be treated (the ’Golden Rule’, Wattles, 1997; Popper, 2012); but also as posing a challenge to balance one’s own values with those of others (captured in the Jewish sage Hillel’s maxim, “If I am not for myself, who will be for me? But if I am only for myself, who am I?”). Different moral principles (as suggested in the opening quote from Nagel) can come into conflict.

For instance one might be forced to choose between helping the lives of many anonymous strangers versus helping a single loved one. Quantitative weighting of the various principles is a natural way to resolve these conflicts while capturing ambiguity.

On this view, moral learning is the process of learning how to value (or "weight") the utilities of different groups of people. Young children and even infants make inferences about socially positive actions and people that are consistent with inference over recursive utility functions: being helpful can be understood as one agent taking another agent's utility function into account in their own decision (Ullman et al., 2009; Kiley Hamlin et al., 2013). Young children also show evidence of weighting the utilities of different individuals, depending on their group membership and social behaviors, in ways that strongly suggest they are guided by abstract moral principles or an intuitive moral theory (Rhodes & Wellman, 2016; Rhodes, 2012; Rhodes & Chalik, 2013; Powell & Spelke, 2013; Hamlin, Mahajan, Liberman, & Wynn, 2013; Hamlin, 2013; Barragan & Dweck, 2014; Kohlberg, 1981; Shaw & Olson, 2012; Smetana, 2006). On the other hand, children do not weight and compose those principles together in a way consistent with their culture until later in development (Hook & Cook, 1979; Sigelman & Waitzman, 1991; House et al., 2013). Different cultures or subcultures might weight these principles in different ways, generating different moral theories (Schäfer, Haun, & Tomasello, 2015; Graham, Meindl, Beall, Johnson, & Zhang, 2016) and posing an inferential challenge for learners who cannot be pre-programmed with a single set of weights. But under this view, it would be part of the human universal core of morality – and not something that needs to be inferred – to have the capacity and inclination to assign non-zero weight to the welfare of others.

The second key component of our model is an approach to inferring others' abstract moral theories from their specific moral behaviors, via hierarchical Bayesian inference. Our analysis of moral learning draws on an analogy to other problems of learning abstract knowledge from observational data, such as learning the meanings of words or the rules of grammar in natural language (Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum et al., 2011). Theorists have long recognized that moral learning, like language learning, confronts children with a challenge known as the "poverty of the stimulus" (Chomsky, 1980; Mikhail, 2006, 2011): the gap between the data available to the learner (sparse and noisy observations of interactions between specific individuals) and what is learned (abstract principles that allow children to generalize, supporting moral tradeoffs in novel situations and for new individuals). More specifically in our framework for moral learning, the challenge of explaining how children learn cultural appropriate weights for different groups of people may be analogous to the challenge of explaining linguistic diversity, and may yield to similar solutions, such as the frameworks of "principles and parameters" (Chomsky, 1981; M. C. Baker, 2002) or Optimality Theory (Prince & Smolensky, 2008). In these approaches, language acquisition is either the process of setting the parameters of innate grammatical principles, or the ranking (qualitatively or quantitatively) of which innate grammatical constraints must be taken into account. Our framework suggests a parallel approach to moral learning and the cultural diversity of moral systems.

So then how do we learn so much from so little? A hierarchical Bayesian approach has had much recent success in explaining how abstract knowledge can guide learning and inference from sparse data as well as how that abstract knowledge itself can be acquired (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Tenenbaum et al., 2011; Xu & Tenenbaum, 2007; Perfors, Tenenbaum, & Regier, 2011; Ayars & Nichols, 2017; Nichols, Kumar, Lopez, Ayars, & Chan, 2016), and fits naturally with the idea that learners are trying to estimate a set of weighted moral principles. By inferring the underlying weighting of principles that dictate how the utility of different agents are composed, a Bayesian learner can make generalizable predictions in new situations that involve different players, different numbers of players, different choices, etc (Heider, 1958; Malle, Moses, & Baldwin, 2001; C. L. Baker et al., 2009; Ullman et al., 2009; Kleiman-Weiner et al., 2015; Goodman, Tenenbaum, & Gerstenberg, 2015; Jara-Ettinger et al., 2016). These hierarchical models allow for a few indeterminate observations from disparate contexts to be pooled together, boosting learning in all contexts (Kemp, Perfors, & Tenenbaum, 2007).

The third key component of our model addresses the dynamics of moral learning. That is, even once children have inferred the moral values of others, when and how are learners motivated to acquire or change their own values? A parallel question at the societal level is what might control the dynamics of moral change across generations. Again we are inspired by analogous suggestions in the computational dynamics of language learning, which has suggested a close relationship between the process of language learning and the dynamics of language change (Christiansen & Kirby, 2003; K. Smith, Kirby, & Brighton, 2003; Niyogi, 2006; Kirby, Cornish, & Smith, 2008; Griffiths & Kalish, 2007; Chater, Reali, & Christiansen, 2009). Children are seen as the main locus of language change, and the mechanisms of language learning within generations become the mechanisms of language change across generations. In that spirit we also consider mechanisms of moral learning that can account for the dynamics of learning both in individuals and at the societal level, for how morals change both within and across generations.

We propose that learners change their own abstract moral values in accordance with two motivations (or meta-values). The first, external alignment, expresses the idea that learners will internalize the values of the people they value, aligning their moral theory to those that they care about (Hurka, 2003; Magid & Schulz, 2017). This mechanism could be associated with a child acquiring a moral theory from a caregiver. It is in some ways analogous to previous proposals for the origins of prosocial behavior based on behavioral imitation or copying behaviors, a mechanism proposed in economics and evolutionary biology both as a primary mechanism of social learning within generations, as well as a mechanism of how prosocial behaviors (including altruism and other "proto-moral" concepts) can evolve across generations (Trivers, 1971; M. A. Nowak, 2006; Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009; Delton, Krasnow, Cosmides, & Tooby, 2011; Rand & Nowak, 2013; Henrich & Gil-White, 2001; Richerson & Boyd, 2008). Pure behavioral imitation is not sufficient to drive learning of the abstract principles and weights that comprise our moral theories (Nook, Ong, Morelli, Mitchell, & Zaki, 2016), but the mechanism of external alignment represents a similar idea at the level of abstract principles and weights.

External alignment alone, however, is not sufficient to explain moral learning or the most compelling aspects of moral change. Across generations, external alignment tends to diffusion and averaging of individuals' moral weights across a society. It cannot explain where new moral ideas come from in a society, or how the individuals in a group can collectively come to value people that few or none of their progenitors valued. Such moral progress is possible. For instance, over the past hundred years there has been significant moral change in racial attitudes and the rights of women in some cultures (Singer, 1981;

Pinker, 2011). What can account for these shifts, or even more strikingly, for the rapid change of moral values in a few or even a single generation as seen recently in attitudes towards same-sex marriage (Baunach, 2011, 2012; Broockman & Kalla, 2016)?

One recent proposal for a cognitive mechanism that underlies moral change is moral consistency reasoning (R. Campbell & Kumar, 2012). R. Campbell and Kumar (2012) describe a dual process account of how deliberative moral judgments are adjusted under pressure from conflicting intuitive responses to analogous moral situations or dilemmas. Inspired by this account, we suggest a second meta-value, internal alignment, where learners try to reduce the inconsistency between their moral theory and their attitudes towards specific individuals. For example, if a learner with parochial values develops feelings for one out-group member, the value she places on all members of that group may shift. During internal alignment, learners adjust their weights over the moral principles to be consistent with feelings about other agents from sources (deliberative and emotional) such as: empathy (A. Smith, 1759; D. Pizarro, 2000; Hoffman, 2001), imagination and stories (Bloom, 2010), analogical reasoning (Keasey, 1973; R. Campbell & Kumar, 2012), love, or involved contact (even imagined or vicarious) (Allport, 1954; S. C. Wright, Aron, McLaughlin-Volpe, & Ropp, 1997; Pettigrew & Tropp, 2006; Shook & Fazio, 2008; Paluck & Green, 2009; Crisp & Turner, 2009). If a learner values a specific agent in a way that is not explained by the moral theory, she will adjust her moral theory to appropriately value that person resolving the inconsistency. Since moral theories are abstract with respect to a particular individual, that realignment may result in rapidly expanding the types of agents that the learner values.

We now present this model of moral learning in full detail. We will describe in turn how moral theories are represented, how they can be inferred from sparse data and how moral acquisition proceeds through meta-values. Finally we turn to the dynamics of moral change and investigate when moral theories will change rapidly and when such change will be slow or nonexistent.

3.1 Representing Moral Theories

The first challenge for moral learners, in our framework, is to represent moral theories for making welfare trade-offs across an infinitude of situations. We start by considering a simplified decision-making environment for this purpose. Let N be a set of agents indexed by i, S be a set of states and A_s be the set of actions available in each state s. The probability of reaching outcome s' upon taking action a in state s is $P(s'|a,s)$ which describes how actions affect outcomes in the world. Let $R_i(s)$ map outcomes to a real number that specifies the welfare agent i intrinsically experiences in state s. Again, welfare can go beyond pleasure and pain but this function maps all of the "base goods" and "base evils" into a single dimensional measurement of overall welfare. Different states may be valued differently by different agents or may vary across different contexts. Thus $R_i(s)$ allows for quantitative assessment of the moral value of a state for a particular agent. In this work, each state presents an agent with a set of choices that can affect its own welfare and the welfare of other agents. Appendix 3.5 gives the details for the decisions studied in this work.

We define moral theories in terms of recursive utility functions which build on $R(s)$ – the welfare obtained by each agent. By defining moral theories in the same units as choice (utility) these moral theories can be easily integrated into a general decision making framework. The level-0 moral theory describes an agent who only cares about the quantity of welfare that she personally receives herself:

$$U_i^0(s) = R_i(s)$$

Thus agents acting consistent with a level-0 moral theory will always choose actions that maximally benefit their own welfare regardless of the effect of that action on the welfare of others. For instance, when faced with the decision to give up a small amount of welfare to provide a large benefit to someone else or doing nothing, an agent acting under a level-0 moral theory would prefer to do nothing. Furthermore, this level-0 theory also has no way of trading off the welfare of other people.

We now build on this selfish agent to account for richer social preferences. In Hurka (2003) the space of values is expanded to include virtue and vices by recursively valuing attitudes towards the “base goods” and “base evils” (e.g., the virtue benevolence as “loving good”). We borrow this idea and extend it to recursively valuing other people to explain social preferences. We define a level-1 moral theory recursively in terms of the level-0 moral theory:

$$U_i^1(s) = (1-\gamma_i)U_i^0(s) + \gamma_i \sum_{\substack{j \in N \\ j \neq i}} \alpha_{i,j} U_j^0(s) \tag{3.1}$$

where $\gamma \in [0,1]$ trades off how much an agent with a level-1 moral theory values their own level-0 utility compared to the level-0 utility of others. When $\gamma_i = 0.5$ agents weigh their own utility equally with the utility of the other agents, when $\gamma_i = 0$ they only care about themselves and when $\gamma_i \geq 0.5$ they value others more than themselves. Generally speaking, γ_i determines the degree to which agent i is prosocial. Each $\alpha_{i,j} \in [0,1]$ is the weight agent i places on the utility of agent j. Depending on the relative value of each $\alpha_{i,j}$, an agent acting under a level-1 moral theory will value some agents more than others. If $\alpha_{i,j} > \alpha_{i,k}$ then agent i cares more about the utility of agent j than the utility of agent k. Since these recursive utilities eventually ground in the welfare of the individual agents, the settings of these parameters specify an entire space of moral theories where the goals and welfare of other agents are treated as ends. Moral theories of this form share similarities to the social preferences used in behavioral game theory but extend those models to consider how different agents might be differentially valued (C. Camerer, 2003). We consider further extensions to these representations in Appendix 3.6.

Having specified a representation for moral theories in terms of recursive utility functions, we consider agents who act consistently with these moral theories using the standard probabilistic decision-making tools. Since our moral theories were constructed from utility functions they can easily be mapped from values into actions and judgments. Since actions can lead to different outcomes probabilistically, decision making and judgment approximately follow from the expected utility of an action:

$$EU(a,s) = \sum_{s'} U(s')P(s'|a,s) \tag{3.2}$$

From expected utility, action selection is defined probabilistically under the Luce-choice decision rule which reflects utility maximization when there is uncertainty about the exact utility value (Luce, 1959):

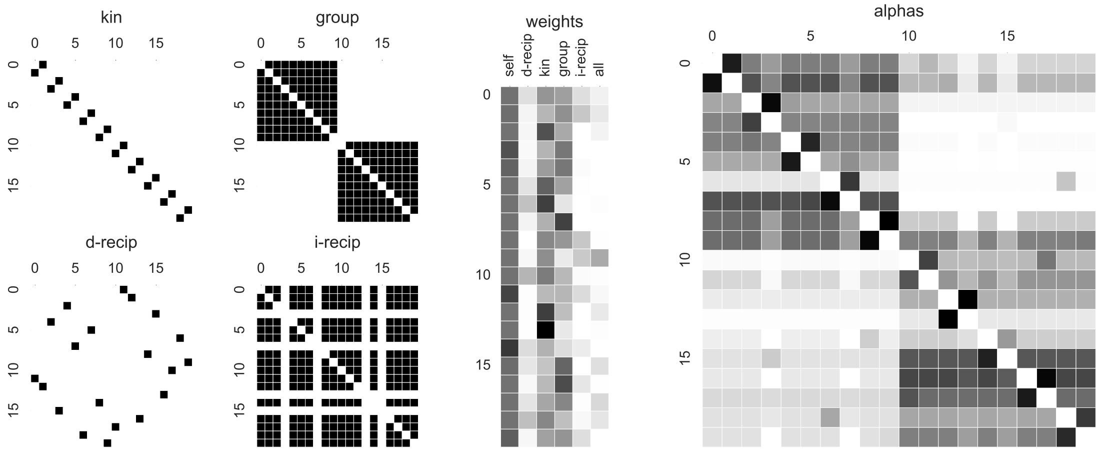
$$P(a|s) = \frac{\exp(\beta EU(a,s))}{\sum_{a' \in A_s} \exp(\beta EU(a',s))} \tag{3.3}$$

In the limit $\beta \to 0$ the decision maker chooses randomly, while in the limit $\beta \to \infty$ the decision maker will always choose the highest utility action.

Thus far we have specified the machinery for a moral agent where the $\alpha_{i,j}$ define how each agent values the others. However, each $\alpha_{i,j}$ describe how a specific person should be valued rather than how to trade-off abstract principles. Without abstract principles an agent would need to specify a new $\alpha_{i,j}$ for every possible individual. Instead, we propose that values over specific people should be determined by more abstract relationships, captured in abstract moral principles: through these principles an agent can deduce how to value anyone.

While there are many ways of specifying the structure of the moral principles in theory, in this work we consider six kinds of relationship that carry moral obligation: (a) self, (b) kin, (c) in-group, (d) all-people, (e) direct-reciprocity, and (f) indirect-reciprocity. For instance, a kin relation might provide a moral reason for helping a loved one rather than an anonymous person. In-group might capture any shared group affiliation that a culture or context defines as morally relevant: gender, ethnicity, nationality, religion, and so on. Direct reciprocity here captures moral obligations to specific known and cooperative individuals (e.g. a person's particular friends and neighbors). Indirect reciprocity captures the moral obligations to members of a broader cooperative community (friends of friends, employees of the same organization). Throughout this work we will assume that agents are not planning about the future-repercussions of their actions and that reputational or direct-reciprocal advantages and disadvantages will be captured by one of the two reciprocity principles.

Each of these principles expresses a simplified type of relationship between agents and gives a reason for the way a decision-maker might act towards a particular person. Since any given dyad may have multiple relations (e.g., a dyad where both individuals are from



(a) (b) (c)

Figure 3-1: A population of 20 agents used throughout this work. (a) Black squares indicate the presence of a relationship for each of the four principles shown. (b) The relative weights on each of the six principles for all 20 agents where each row is the weighting of principles of a single agent. Darker values correspond to a higher weight. (c) The $\alpha_{i,j}$ parameters implied by the weights and relationships. The darker the cell the more weight that the agent indexed by the cell's row puts on the agent indexed by the cell's column.

the same in-group but also have a direct reciprocity relationship), each principle is associated with a corresponding weight that quantitatively describes how that principle is traded-off against others. Neural evidence of these principles has been detected in cortical and limbic brain circuits (Rilling et al., 2002; Krienen, Tu, & Buckner, 2010; Watanabe et al., 2014) and there is some evidence that the relative strength of these circuits can provide motivation for certain types of altruistic behavior (Hein, Morishima, Leiberg, Sul, & Fehr, 2016).

Formally, let $P = \{\texttt{kin}, \texttt{group}, \ldots\}$ be the set of moral principles. Then for each principle there is a function $f^p(i,j)$ over pairs of agents that returns 1 if the relationship between i and j falls under principle p and 0 otherwise. Specifically, $f^{\texttt{kin}}(i,j) = 1$ if i and j are kin, $f^{\texttt{group}}(i,j) = 1$ if i and j are in the same in-group and $f^{\texttt{all}}(i,j) = 1$ for all $i \neq j$. $f^{\texttt{self}}(i,j) = 1$ for all $i = j$. The $f^{\texttt{d-recip}}(i,j) = 1$ if i and j have a reciprocal relationship and $f^{\texttt{i-recip}}(i,j) = 1$ if both i and j are in the cooperative group (M. A. Nowak & Sigmund, 2005). We assume all principles are symmetric so $f(i,j) = f(j,i)$ and that the relationships are binary (present or absent). These principles encode abstract knowledge about relationships between agents rather than knowledge about specific agents.

Figure 3-1a visualizes these relationships for a population of 20 agents. In this population each agent has a single kin relationship and belongs to one of two groups. Note that the direct-reciprocity relationships are sparse. Since direct-reciprocity is a reciprocal relationship between two agents, it is not necessarily transitive. Just because i has a reciprocal relationship with j and j has a reciprocal relationship with k, it does not necessarily follow that i and k will also have a reciprocal relationship. In contrast, indirect-reciprocity denotes membership in a cooperative or trustworthy group (M. A. Nowak & Sigmund, 2005). These relationships are based on group identity such that everyone in the cooperative group has an indirect-reciprocity relationship with everyone else in the cooperative group. Hence these relationships satisfy transitivity. Unlike previous formal models of reciprocity that were defined in terms of specific behaviors in specific situations, such as Tit-for-Tat in the prisoners dilemma (Axelrod, 1985; M. A. Nowak, 2006; Rand & Nowak, 2013), our principles of reciprocity are implemented in agents who can reciprocally value the utility of each other. These more abstract concepts of reciprocity (direct and indirect) lead to moral judgments and actions that generalize robustly across different situations and contexts.

These principles are then weighted so they can be quantitatively traded off. Let W_i be the weights that agent i places over the moral principles. Each $w_i^p \in W_i$ is the weight that agent i places on principle p. For self valuation, let $\gamma_i = 1 - w_i^{self}$. We now rewrite the $\alpha_{i,j}$ of equation (3.1) as a function of weights over moral principles:

$$\alpha_{i,j}(W_i) = \phi_{i,j} + \sum_{p \in P} w_i^p \cdot f^p(i,j) \tag{3.4}$$

Unlike $\alpha_{i,j}$ which define *who* each agent values, the W_i define *what* each agents values. Who each agent values ($\alpha_{i,j}$) can be derived using equation (3.4) from what that agent values i.e., their weights over principles W. We introduce an additional source of valuation $\phi_{i,j}$ which stands in for other factors outside of the moral principles that describe how i values j. Figure 3-1c shows the $\alpha_{i,j}$ derived from the weights and relations of Figure 3-1.

3.2 Inferring Moral Theories

Above we described how moral theories, expressed as weights or values placed on abstract relationships and then composed in a recursive utility calculus, can be used during moral decision making and judgment. That is, we described the forward model, in which moral decision makers can use their moral theories to choose actions and judgments in any context. The second challenge for moral learners is to infer how others weight the abstract moral principles from sparse and noisy observations. In the same way that rational actors reveal information about their beliefs and desires through their behavior, moral agents reveal information about their moral theory through their behavior and judgments.

Expressing the intuitive theory in terms of principles over abstract categories helps to make learning tractable. Rather than inferring the value of each $\alpha_{i,j}$ independently, a learner only needs to determine how to weigh a relatively smaller set of moral principles. It is the abstractness of the principles that enables generalization and rapid learning under the "poverty of the stimulus" (Kemp et al., 2007). If a learner observes that a particular agent weights kin highly, and a new person is introduced who is also related to that agent, the learner will already have a good idea of how this new relative will be valued. Knowledge of abstract weights can often be acquired faster than knowledge of particulars, which is sometimes called "the blessing of abstraction" or "learning to learn" (Kemp et al., 2007; Kemp, Goodman, & Tenenbaum, 2010; Goodman, Ullman, & Tenenbaum, 2011). This is the power of hierarchical modeling.

Learning abstract principles also clarifies the intuitive idea that people in a given culture or in-group will agree more about the relative value of abstract moral principles than about the relative value of specific people. For instance, people in a specific culture might each highly value their own siblings but not the siblings of others. Thus we want to model the way that these theories will be learned at the level of principles not at the level of individuals. Moral principles explain how moral learners can go beyond the data and infer hierarchical abstract theories from behavioral data.

Note that we assume that `self`, `kin`, `in-group` and `all-people` are observable to the learner i.e., the learner knows which agents are kin and which belong to a common in-group

(DeBruine, 2002; Lieberman, Tooby, & Cosmides, 2007). However, when observing interactions between third parties, relationships based on reciprocity (`direct` and `indirect`) are not directly observable by the learner and need to be inferred from behavior. Sensitivity to these principles could be innate but could also be learned from a sufficiently rich hypothesis space or grammar of abstract knowledge (Goodman et al., 2011; Tenenbaum et al., 2011).

We can now formally state the challenge of inferring a moral theory. Let T be the number of observations made by the learners. Most of the specific choices we make for the hierarchical model are not essential for our cognitive argument, but are useful to facilitate implementation and simulation. While we are committed to a hierarchical structure in general, the specific mathematical forms of the model (e.g., the choice of priors) are at most provisional commitments; they are chosen to be reasonable, but there are many possible alternatives which future work could investigate. Each observation (a_i, s) is information about the choice a_i made by agent i from the choices available in state s. For a learner to infer the moral theories of others, she needs to infer the weights over the moral principles conditional on these observations, $P(W|(a_i^0, s^0), \ldots, (a_i^T, s^T))$. This conditional inference follows from Bayes' rule:

$$\begin{aligned} &P(W_i|(a_i^0, s^0), \ldots, (a_i^T, s^T)) \propto \\ &\quad \sum_{f^{\texttt{d-recip}}} \sum_{f^{\texttt{i-recip}}} P(a_i^0, \ldots, a_i^T | s^0, \ldots, s^T, W_i, f^{\texttt{d-recip}}, f^{\texttt{i-recip}}) P(W_i) P(f^{\texttt{i-recip}}) P(f^{\texttt{d-recip}}) \end{aligned} \tag{3.5}$$

where the likelihood $P(a_i^0, \ldots, a_i^T | s^0, \ldots, s^T, W_i, f^{\texttt{d-recip}}, f^{\texttt{i-recip}})$ is probabilistic rational action as shown in equation (3.3) with the $\alpha_{i,j}$ set by the weights over moral principles as shown in equation (3.4). To complete this hierarchical account of inference, we need to specify priors over the unobserved principles direct-reciprocity and indirect-reciprocity and over the weights themselves.

Since direct-reciprocity relationships are sparse and non-transitive we put an exponential prior over each possible reciprocal relationship (B. Lake & Tenenbaum, 2010):

$$P(f^{\texttt{d-recip}}) = \prod_{i \in N} \prod_{\substack{j \in N \\ j \neq i}} \lambda \exp(\lambda f^{\texttt{d-recip}}(i,j))$$

This prior generally favors a small number of direct-reciprocity relationships when observations are ambiguous. The higher the value of λ, the more unlikely these relationships.

Indirect-reciprocity relationships are an inference over the group rather than individual dyadic relationships. Each agent is either in the "cooperating group" or not, and only when both are in the cooperating group will they value each other under the indirect-reciprocity relationship. Here C is the "cooperating group":

$$P(f^{\texttt{i-recip}}) = \prod_{i \in N} p^{\mathbb{1}(i \in C)}(1-p)^{\mathbb{1}(i \notin C)}$$

with p as the prior probability of an agent being in the "cooperating group".

Having specified priors for the two unobserved reciprocity principles, we now describe how learning abstract knowledge about how moral theories are shared within groups allows learners to rapidly generalize their knowledge. We define a generative model over the possible ways the principles could be weighted $P(W)$. The simplest model might treat each individual's weights as generated independently from a common prior, reflecting a belief in some "universal human nature". Here we consider a more structured model in which learners believe that individual's weights are drawn from a distribution specific to their group. This represents group moral norms that themselves should be inferred in addition to the weights of individuals. Specifically we assume that the weights of each individual W_i are drawn from a Gaussian distribution parameterized by the average weighting of principles in that individual's group g:

$$W_i \sim \text{Normal}(W^g_{\texttt{norm}}, \Sigma^g)$$

where $W^g_{\texttt{norm}}$ is the average weighting of principles in i's group and Σ^g is how these weights covary in different individuals of a group. After sampling, the weights are normalized

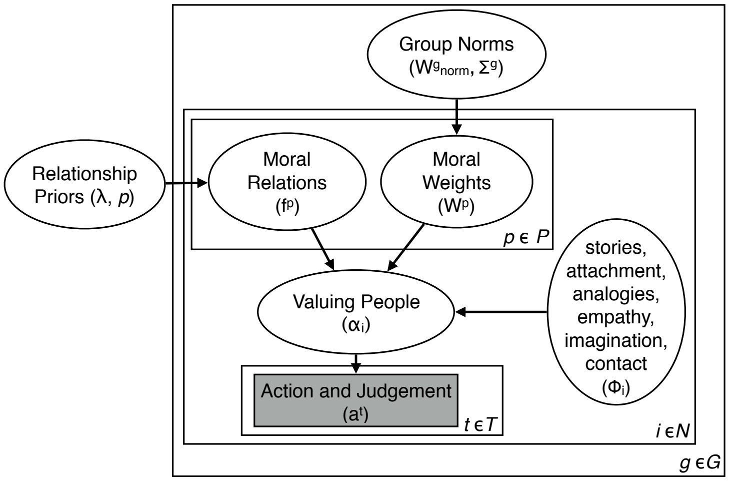

Group Norms
(W^g_norm, Σ^g)
Relationship Priors (λ, p)
Moral Relations (f^p)
Moral Weights (W^p)
p ∈ P
stories, attachment, analogies, empathy, imagination, contact (Φ_i)
Valuing People (α_i)
Action and Judgement (a^t)
t ∈ T
i ∈ N
g ∈ G

Figure 3-2: Hierarchical probabilistic model for inferring latent moral theories from sparse behavior. T is the number of actions and judgments observed, N are the agents, P are moral principles and G are the groups. Actions and judgments are observed (shaded in gray).

so that they are positive and sum to one. The higher the values in Σ^g the more variance there will be in how agents weight the principles. The correlation between the weights of the agents is visible in Figure 3-1b. Importantly, a learner does not know the $W^g_{\texttt{norm}}$ for each group g in advance. The group average $W^g_{\texttt{norm}}$ must be inferred jointly with the W_i of each agent. Thus while each person has a unique set of weights over moral principles, those weights are statistically correlated with the weights of others in their group since they are drawn from the same latent distribution. In this work we consider only diagonal Σ^g for simplicity which do not model how principles themselves might be correlated. For instance, in some society agents that highly weight the `kin` principle may also highly weight the `group` principle highly. These correlations could be captured by allowing for covariance in Σ^g. The full hierarchical model is shown schematically in Figure 3-2.

Assuming this structure for $P(W)$ is just one possible way to add hierarchical structure to the inference of moral theories. Instead of inferring a different $W^g_{\texttt{norm}}$ for each group, the learner could infer a single $W_{\texttt{norm}}$ for all agents which would imply that the learner assumes moral theories do not systematically vary across groups. Furthermore, the $W^g_{\texttt{norm}}$ themselves could vary in a systematic way according to a universal prior. For instance while one might expect all groups to value `kin` highly but show significant diversity in how much they care about `group`. We did not vary Σ^g in this work but one can imagine a learner inferring that some groups have more within group moral diversity than others which would be captured by joint inference over this parameter.

We now empirically investigate inference in this model via a set of simulations. One of the key reasons to use utility functions to represent moral theories is that our learner can learn from observing different kinds of decisions and judgments in different contexts: they do not need to see many examples of the same decision, as in classic reinforcement learning and learning-in-games approaches (Fudenberg & Levine, 1998). In our simulations, observations of judgments and decisions took two forms: either the actor traded off her own welfare for that of another person or the actor traded off the welfare of one agent for the welfare of another. Within these two types, each observed decision was unique: The actors involved were unique to that interaction, and the quantities of welfare to be traded off were sampled independently from a probability distribution of characteristic gains and losses. See Appendix 3.5 for the specific details of the judgments and decisions used as observations.

Another feature of our simulations is that learners' observations of behavior are highly biased toward their kin and in-group (Brewer & Kramer, 1985). This makes learning more difficult since most of the observed data is biased towards just a few agents but the learner needs to infer weights and principles that apply to all agents. Figure 3-3 shows an example of the inference for $P(W|(a_i^0, s^0), \ldots, (a_i^T, s^T))$ and the marginalized reciprocity relationships $P(f^{\texttt{d-recip}}, f^{\texttt{i-recip}}|(a_i^0, s^0), \ldots, (a_i^T, s^T))$. As the learner observes more data, the inferences become more and more accurate. However even with just a few observations, hierarchical Bayesian inference leverages both the abstract principles and the hierarchical prior over the weights of groups to rapidly approximate the moral theories of others.

3.3 Moral Learning as Value Alignment

Having described how rich moral theories can be represented and efficiently inferred from the behavior of others, we now turn to moral learning itself. Specifically, how do moral learners set their own weights over principles? We propose that moral learners have meta-values, or preferences over moral theories themselves. Moral learning is then the process of

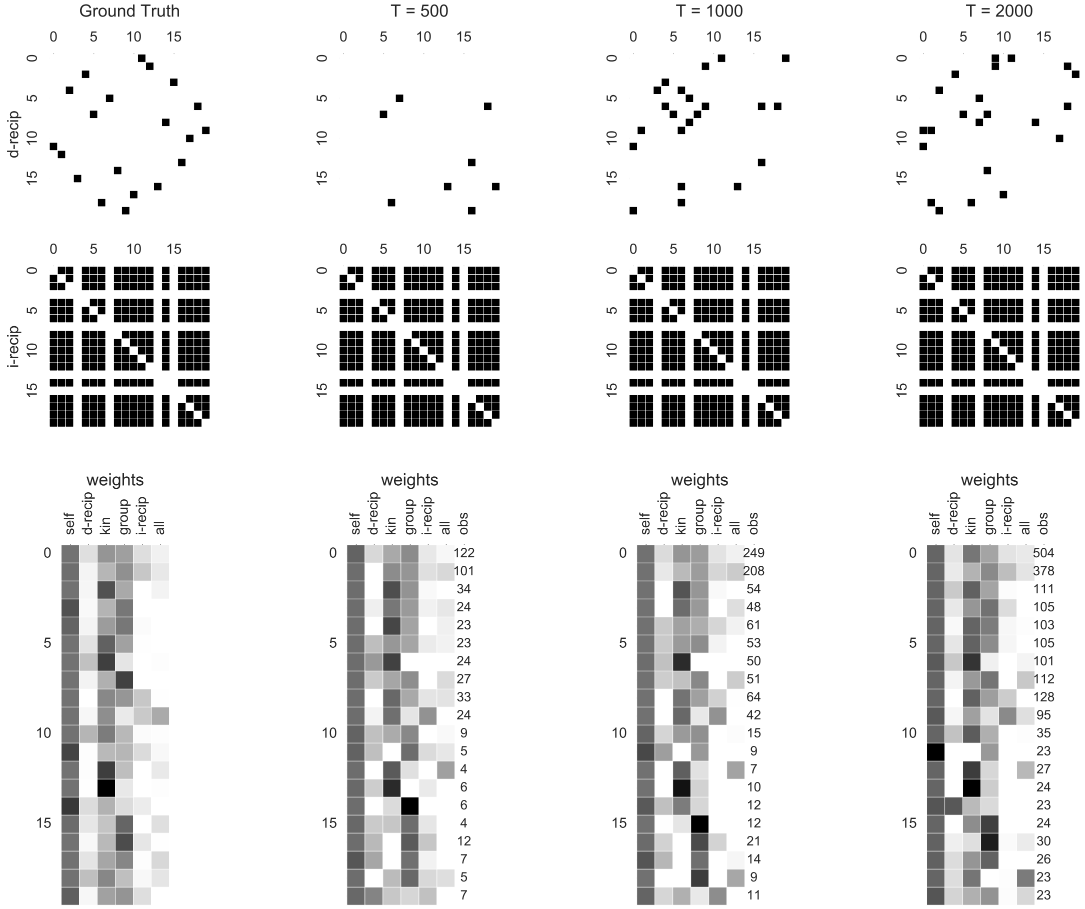


Figure 3-3: Maximum a posteriori (MAP) estimate of beliefs from a learner observing behavior from the society shown in Figure 3-1 under increasing observations ($T = \{500, 1000, 2000\}$). This learner is biased towards observing the behavior of agents 0 and 1. (top) Samples of the graph inference for the two reciprocity principles. The indirect-reciprocity relationships are inferred rapidly while direct-reciprocity is slower and more error prone because of its sparsity. (bottom) The weights inferred by the learner for each of the other agents. The learner rapidly infers the moral theories of its kin (rows 0-1) and in-group (rows 0-9) but has significant uncertainty about the moral theories of agents in its out-group (rows 10-19). The "obs" column is the number of times the learner observed that agent make a moral decision. Note that the vast majority of the observations come from kin and the in-group. See Appendix 3.5 for the details of the inference.

aligning a moral theory with these meta-values. We propose two types of meta-values and study specific instantiations of them. The first, external alignment, instantiates a form of social learning where learners try to align their weights over principles as close as possible to the weights of those that they value. The second, internal alignment, is a meta-value for a moral theory which is consistent with the learner's attachments and feelings. We formalize these meta-values for moral theory alignment and show that they can provide insights into understanding the dynamics of moral change.

3.3.1 External Alignment: Learning from others

External alignment is a form of cultural or social learning. We explicitly depart from the type of social learning commonly used in evolutionary models of game theory which depend on behavioral imitation or learning by reward reinforcement (M. A. Nowak, 2006; Richerson & Boyd, 2008; Rand & Nowak, 2013). Instead, we propose that learners acquire a moral theory by internalizing the abstract principles used by others. Since we have already described how a learner can infer the moral theories held by other agents, we now describe how a learner decides *who* to learn from (Henrich & Gil-White, 2001; Richerson & Boyd, 2008; Rendell et al., 2010, 2011; Frith & Frith, 2012; Heyes, 2016).

We propose that a learner L sets their moral theory to be close to the moral theories of those whom they value. We express this meta-value as a utility function that the learner is trying to maximize with respect to their weights over principles. The utility function measures how similar the learner's weights are with the weights of the people that the learner values. Since who the learner values is determined in part by their weights, there is an implicit dependence on their current weights, $\hat{w}_L$:

$$U_{\texttt{external}}(w_L|\hat{w}_L) = -\sum_{i\in N} \alpha_{L,i}(\hat{w}_L) \sum_{p\in P} (w_L^p - w_i^p)^2. \tag{3.6}$$

This utility function has two nested sums. The inner sum over principles p is the sum of squares difference between the moral weighting of the learner and of agent i for each principle p. Maximum a posteriori (MAP) estimates were used for the inferred weights w_i of the other agents. The outer sum over agents i sums that squared difference weighted by how much the learner values each agent i, $\alpha_{L,i}(\hat{w}_L)$, given their current weights $\hat{w}_L$. Recall that $\alpha_{i,j}(\hat{w}_L)$ is composed of two terms: a sum over the moral principles as well as an additional ϕ term which can contain other feelings and attachments that are not characterized by the moral principles as shown in equation (3.4). We propose that a learner may have some special attachments or feelings towards certain people. Particularly in the case of theory acquisition we consider a primitive attachment towards a caregiver which results in a learner having a high ϕ directed towards that person (Bandura & McDonald, 1963; Cowan, Longer, Heavenrich, & Nathanson, 1969; Hoffman, 1975; Govrin, 2014). It is interesting to note that this utility function has a similar structural appearance to the utility function of the moral decision maker shown in equation (3.1). If we imagine that agents have a preference that others share their values, then a learner is increasing the utility of the people she values by matching her weights to their weights.

To see how the internalization of the values of others might work dynamically, consider a learner with a single primitive attachment to person i so that $\phi_{L,i} > 0$. By valuing person i, the learner will need to bring her weighting of moral principles in line with i's weighting to minimize $\sum_{p \in P}(w_L^p - w_i^p)^2$. But by bringing her values (as characterized by her weights over moral principles) inline with those of agent i, she will start to value other agents as well. This process can repeat, with the updated weights w_L becoming the old weights $\hat{w}_L$. For instance, if L and j are in the same in-group and i (L's caregiver) weights in-group highly then when L brings her values in line with i, she will also start to value j since $w_L^{\text{group}} > 0$ implies $\alpha_{L,j}(w_L) > 0$. But since $\alpha_{L,j}(w_L) > 0$, the learner will also try to bring her values inline with the values of j (although to a lessor degree than i). Through this mechanism, a learner who starts off valuing only a single specific person (e.g., their caregiver) will initially internalize just that person's values. But adopting that person's values may entail valuing other agents and the learner will recursively average the weights of those agents into her own. The model makes the non-trivial claim that the $\alpha_{i,j}$ parameters perform a dual role: they are both the target of inference when learning from the behavior of others, and they also drive the acquisition of the moral knowledge of others.

We empirically investigate the dynamics of external alignment in the previous society of agents (Figure 3-1). Each of the 20 agents act as a caregiver (with a corresponding primitive attachment) to a single learner. Figure 3-4 (top) shows the equilibrium weights of the 20 learners. The weights that each learner acquires are a combination of what they infer the weights of their caregiver to be and the inferred weights of the other agents. The extent to which the weights of other agents are ultimately mixed in with the caregivers' weights is controlled by the ϕ on the learners caregiver. As Figure 3-4 shows, when this ϕ is high, the

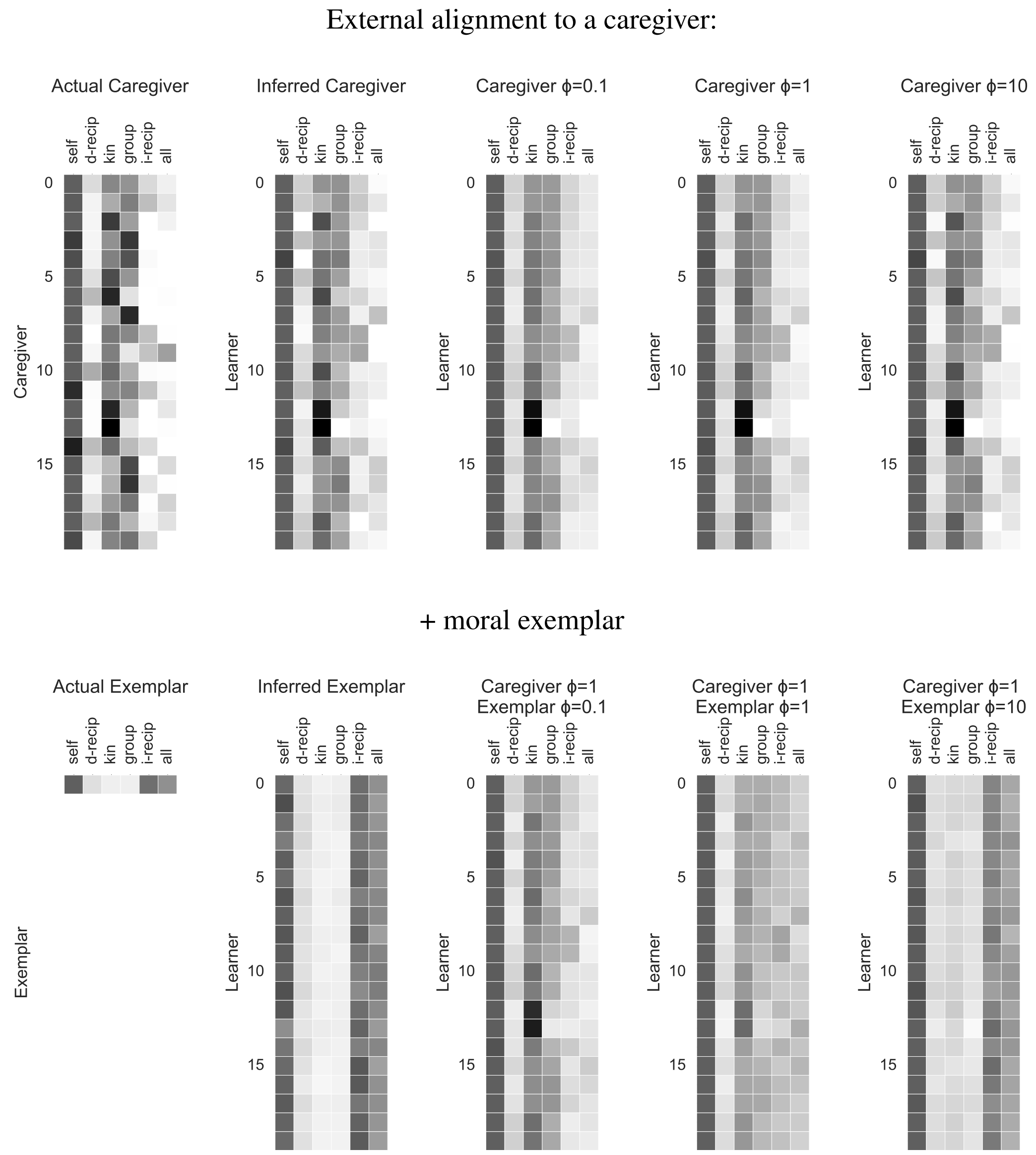


Figure 3-4: External alignment with caregivers and moral exemplars. The "Actual" columns shows the actual weights for the caregivers of each of the 20 learners and the moral exemplar. The "Inferred" columns show the weights each learner infers about the weights over principles used by their own caregiver (top) and a highly impartial moral exemplar (bottom). The "Actual" and "Inferred" columns look similar since learners infer weights of others with high fidelity. The following upper columns entitled "Caregiver" show the resulting moral theory actually adopted by each of the 20 learners as a result of the process of external alignment shown in equation (3.6). The different values of ϕ sets the strength of the feelings of the learner towards their caregiver. For low values of ϕ the learners end up valuing many agents and so adopt weights that are similar to the mean weight of their group. As ϕ increases there is less averaging and each agent is more likely to only internalize the weights of their caregiver. The lower columns entitled "Exemplar" show the resulting moral theory when learners internalize both the values of their caregivers and the moral exemplar. As the ϕ on the exemplar increases, learners move from mixing the caregiver with the exemplar to directly inheriting the values of the exemplar.

learner just internalizes the values of their caregiver. When ϕ is low, the learner chooses weights that are somewhat in between her caregiver's weights and the weights of those that the learner ends up valuing.

Beyond this dynamic of acquisition, other ways of setting ϕ can lead to different learning dynamics. For instance, if learners place a high ϕ on agents they aspire to emulate in terms of success or status, the learning dynamic will emulate that of natural selection. This is analogous to the replicator dynamics used in evolutionary game theory but would operate on abstract moral principles rather than behavioral strategies.

In addition to a primitive attachment such as a relationship with a caregiver, one could also emulate moral exemplars. This kind of learning can also drive moral change for better or for worse. Moral figures like Martin Luther King Jr. and Mother Teresa have inspired people not only to copy their specific prosocial actions and behaviors (e.g., protesting for African American civil rights and helping the needy) but to internalize their values of impartial consideration for all. The bottom half of Figure 3-4 shows learners update their weights under the external alignment dynamic when they have feelings for both their own caregiver and a moral exemplar with saint-like impartial values (assigning high weights to the indirect reciprocity and all-people principles). For intermediate values of ϕ towards the exemplar, the learners mix the values of their caregivers with those of the exemplar. For higher values of ϕ towards the exemplar the learners' weights mostly reflect the exemplar. Finally, moral exemplars need not lead to progress. A charismatic dictator or demagogue can inspire others to narrow their moral theory to place more moral weight on one's in-group at the expense of the broader principles.

3.3.2 Internal Alignment: Learning from yourself

While external alignment can account for how values are passed on over time and how new ideas from a moral exemplar can spread, it does not generate new moralities that cannot be described as a combination of moral theories that are already expressed in the society. In a society where everyone only narrowly extends moral rights to others, how can more broad or impartial theories emerge? We now turn to a second possible mechanism for learning,

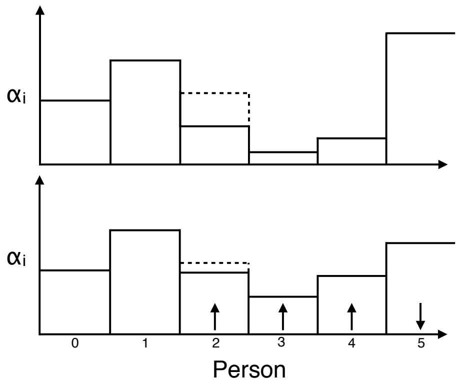


Figure 3-5: Internal moral alignment through inconsistency reduction. (a, top) Schematic of a learner's current moral theory $\hat{w}_L$. The solid line shows the contribution of the moral principles to the $\alpha_{L,i}$ for each of the agents (in arbitrary order). The dotted line is the additional contribution of $\phi_{L,i}$ on the $\alpha_{L,i}$ for a particular agent. (a, bottom) The learner's updated moral theory w_L after internal alignment. This moral theory is adjusted so that the gap between the solid line and dotted line is minimized, which may also affect some of the other $\alpha_{L,i}$ (note the arrows pointing in the direction of the change).

internal alignment, which revises moral theories to generate new values through the reduction of internal inconsistency. Our notion of internal alignment mirrors some aspects of the "reflective equilibrium" style of reasoning that moral philosophers have proposed for reconciling intuition and explicit moral principles (Rawls, 1971; R. Campbell, 2014). We argue that a similar reflective process can also occur within individuals during moral learning and gives insights into how commonsense moral theories change.

We start by supposing that through the course of one's life, one will acquire attachments for various people or even groups of people. These attachments and feelings can be represented through the ϕ vector introduced in the previous section. As mentioned in the introduction, these ϕ values could come from empathy and emotional responses, imagination and stories, morally charged analogical deliberation, love, contact, exposure etc. We do not explicitly model how these diverse mechanisms could lead to the formation or breaking of attachments. Instead we directly manipulate the values of ϕ.

These feelings which also motivate moral valuation of specific individuals (through ϕ) will not necessarily match the weight one's moral theory places on those individuals. This could happen, for instance, when a person with a moral theory that places little weight on anyone outside of their in-group happens to fall in love with an out-group member.

These feelings might affect one's moral theory through a desire for moral consistency: a preference to adopt a moral theory that does not conflict with one's feelings and intuitions (R. Campbell & Kumar, 2012; Horne, Powell, & Hummel, 2015). Said another way, feelings inconsistent with the learner's moral theory could generate an aversive error signal. The learner would then adjust her moral theory in order to reduce the overall magnitude of this signal, aligning her moral theory to be internally consistent with these feelings. This adjustment could be conscious as in moral consistency reasoning (R. Campbell & Kumar, 2012) or unconscious as in cognitive dissonance (Festinger, 1962). Based on this intuition, we propose a second meta-value for choosing a moral theory that captures this reasoning:

$$U_{\texttt{internal}}(w_L|\hat{w}_L) = -\sum_{i \in N}\left[\alpha_{L,i}(\hat{w}_L) - \sum_{p \in P} w_L^p \cdot f^p(L,i)\right]^2. \tag{3.7}$$

This criteria takes the form of a utility function that the learner is trying to maximize with respect to their weights over principles. The utility function measures the difference between how much their moral theory tells them to value each person and how much they actually value that person when their feelings are included. The intuition behind internal alignment is that one wants to find a new moral theory (w_L) that values specific individuals (the sum over P) in a way that is consistent with the way one feels about individuals (the $\alpha_{L,i}$) which includes both moral principles $\sum_{p \in P} \hat{w}_L^p \cdot f^p(L,i)$ and the $\phi_{L,i}$ as shown in equation (3.4). In the case where there are no additional attachments (and hence $\phi_{L,\cdot} = 0$), the two terms will be in alignment and the learner will choose $w_L = \hat{w}_L$ i.e., maintain their original moral theory without change. When these are not in alignment (and hence $\phi_{L,\cdot} \neq 0$), the weights over principles will be adjusted such that they have higher weight on principles that include agents where $\phi_{L,i} > 0$ and lower weight on principles that include agents where $\phi_{L,i} < 0$. A schematic of this process is shown in Figure 3-5.

Consider a father who holds significant homophobic views and treat homosexuals as an out-group. If he discovers that a close friend or even his own child is homosexual, his moral theory is telling him to value that close friend or child much less than he had felt before. In order to align his weights over principles to be consistent with his feelings the father may update his moral theory to place less weight on that in-group relation and

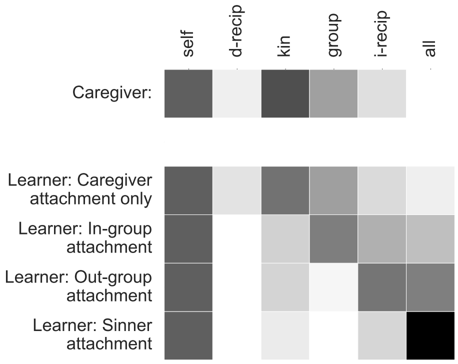


Figure 3-6: Broadening a parochial moral theory through attachments and internal alignment. The caregiver and all other agents have parochial values (shown in the "Caregiver" row) which were inferred by the learner as in Figure 3-3. When the learner only has a primitive attachment for the caregiver (like those shown in Figure 3-4), her moral theory closely reflects the moral theory of the caregiver (shown in the "Caregiver attachment only" row). Each following row shows the resulting moral theory when the learner forms an attachment with an additional individual (with strength $\phi = 1$). When the learner forms an attachment for a person in their in-group their moral values move from kin to in-group. When the learner forms an attachment with someone in their out-group but who is also in the group of indirect-reciprocators, the learner's weights broaden towards indirect-reciprocity. Finally, when the learner forms an attachment with a "sinner," an out-group member who doesn't belong to the group of indirect-reciprocators, the only way to resolve the inconsistency is to highly weight all people.

more weight on the more universal values (all or indirect-reciprocity). Likewise, in the novel "The Adventures of Huckleberry Finn," as Huck develops a bond with Jim, a black runaway slave, his feelings are no longer consistent with the parochial moral weighting he had previously held (where race is the key feature defining groups) and he updates his moral weighting to include Jim, which might also include other black people.

Internal alignment is one way to explain the phenomenon of expanding moral circles, the extension of rights and care to increasingly larger groups of people over time. In our model this corresponds to moving from the narrow values of kin and in-group to more impartial values of indirect-reciprocity and valuing everyone. We first study how this might work at the level of an individual agent. Figure 3-6 shows how a learner's weights over principles move from weighting more parochial to more impartial values in response to new attachments and internal alignment. Crucially and in contrast to external alignment, internal alignment can account for moral change that does not arise from merely copying the values of others. As learners have new experiences, emotional or deliberative, their appreciation of other people may change and the inconsistency generated by those experiences can lead to new moral theories.

Internal alignment is broader than the specific instance studied here and other forms are certainly possible. While we focus on adjusting the weights of the moral theory, the nature of the principle could also be changed. For instance, the father of the homosexual child could also reduce inconsistency by subtyping his in-group/out-group membership criterion such that his child was not excluded (Weber & Crocker, 1983). Another way to reduce inconsistency would be to allow the attachments themselves to change. The father might weaken his feelings for his child. Also note that internal alignment may lead to reducing the moral weight of whole groups. If a learner comes to develop negative feelings for an individual of a certain group (for example after being victimized by crime), that experience may drive them toward a more parochial weighting of principles. Figure 3-7 shows how the narrowing of an impartial theory can occur within a single individual in response to negative attachments and hatred.

In sum, while external alignment leverages primitive relations to learn abstract moral principles, internal alignment modifies moral principles to make them consistent with feelings and relationships. While external alignment can remove disparities between *what* learners weight and what the people they value weight, internal alignment can remove disparities in *whom* the agent values by changing what the learner values. Perhaps the clearest way to appreciate this distinction is to consider the difference between two canonical examples of moral change where these different alignment mechanisms are operative. Consider a learner who “loves a saint” versus a learner who “loves a sinner”. Both situations can lead to moral change, but moral learning by loving a saint follows from external alignment while moral learning by loving a sinner follows from internal alignment. That is, loving the saint will lead to copying the values of the saint, for instance internalizing their weight on the indirect-reciprocity principle as we showed in Figure 3-4 where learners copied from saint-like moral exemplars. But in loving a sinner, the sinner doesn’t have weights that the learner can copy since they presumably conflict with the weights of the other people she values (“love the sinner, hate the sin”). However, internal alignment is still a viable force.

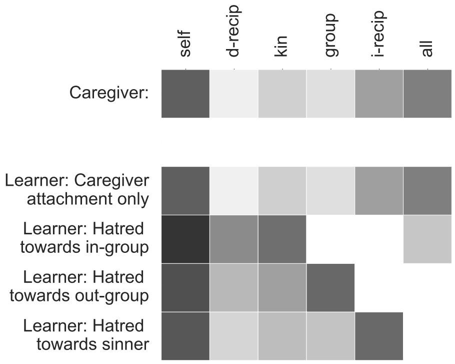
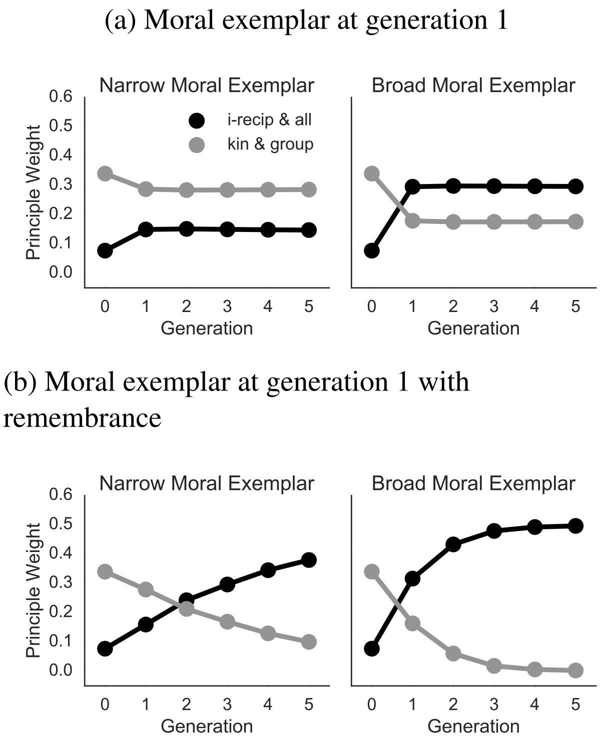


Figure 3-7: Narrowing an impartial moral theory through feelings of hatred and internal alignment. The caregiver and all other agents have impartial values (shown in the "Caregiver" row) so change cannot occur through external alignment. These moral theories were inferred by the learner as in Figure 3-3. When the learner only has a primitive attachment for the caregiver, her moral theory closely reflects the impartial moral theory of the caregiver (shown in the "Caregiver attachment only" row). Each following row shows the resulting moral theory when the learner forms a negative-attachment (hatred) with $\phi = -1$ towards the hated agent. When the learner experiences hatred toward a person in their in-group internal alignment narrows their moral values to just weight kin and direct-reciprocity. When the learner experiences hatred for an out-group member who is also in the indirect-reciprocator group the weights narrow to highly weight the in-group at the expense of all people. Finally, when the learner experiences hatred towards a "sinner," an out-group member who doesn't belong to group of indirect-reciprocators, the inconsistency is resolved by only narrowing away from valuing everyone.

By highly weighting the "all people" principle, the learner can value both the sinner who she loves and the other good people the learner values (as in Figure 3-6). To make these examples concrete, contrast a prejudiced white learner who is inspired to value a moral leader such as Martin Luther King Jr., and a prejudiced white learner who comes to value a specific black person who is not especially virtuous (as Huck Finn did with Jim). The former may copy the impartial values of MLK while the latter may adjust his moral weightings to include that special person in an effort to make his moral theories consistent.

3.3.3 Dynamics of Moral Change

These two learning mechanisms, external and internal alignment, also have implications for the dynamics of moral evolution – how moral values change over generations. In our experiments, for each generation, a new group of learners observe biased samples of be-


(a) Moral exemplar at generation 1

(b) Moral exemplar at generation 1 with remembrance

Figure 3-8: Moral exemplars can rapidly reshape moral theories. When a moral exemplar with impartial values is introduced to parochial minded agents at generation 1 (a), the moral theories immediately adjust. There was a larger shift in moral theories when the moral exemplar was stronger (right) and affected 75% of the agents than when the exemplar was weaker (left) and only affected 25%. However, when the exemplar's influence extends past their lifetime (b) they can continue to reshape moral theories long after that exemplar's direct influence.

havior and judgment from the previous generation, infer the underlying moral theory (as in Figure 3-3) and through value alignment, set the weights on their own moral theory (as in Figure 3-4). This process is iterated for each generation with the learners of the previous generation becoming the actors for the next generation of learners. Using this model of generational learning we are able to formulate and answer questions about how moral learning translates into moral change.

One question, for example, is what leads moral change to persist, and even accelerate across generations. We hypothesize that through external alignment, a moral exemplar might rapidly affect moral values in even a single generation. The more people that are affected by the exemplar (a measure of that exemplar's influence), the greater the shift. Once changed, this shift persists in future generations (Figure 3-8b), but does not continue to grow (and indeed may eventually be lost). Thus, we suggest that the greatest moral change occurs when the exemplar persists across generations in retold stories and memories. As an example, consider the rituals around "sainthood" in which a moral exemplar's good acts are relived and remembered across generations. This persistence allows the exemplar's moral principles to continue to shift moral values long after their original influence (Figure 3-8d).

Another question concerns how rapid moral change can spread through a group even without a specific exemplar (Singer, 1981; Pinker, 2011). For example, how do attachments between specific individuals create systematic change in overall moral norms, via internal alignment?

In our simulations, agents started out with a parochial moral theory which heavily weighted the kin and in-group principles and placed very little weight on the impartial principles of indirect-reciprocity and all people (shown in Figure 3-1). To measure moral change we examined the average weighting of these principles during each generation. In each simulation we varied the fraction of new feelings and attachments ($\phi > 0$) we created in each generation and the distribution of those new attachments across the agents. The proportion of agents ($\rho = 0.05, 0.15, 0.25$) who formed a new attachment towards another agent besides their caregiver varied in each experiment. We analyze the equilibrium of jointly optimizing the external and internal alignment utility functions. Since there are no "saints" in these simulations, internal alignment is necessary for systematic directional change in the average weights of the society.

In the first set of simulations, these attachments were created between agents uniformly at random. Because of uniform sampling, an agent's new attachment is unlikely to be towards someone in their kin group and $\approx 50\%$ likely to be towards someone in their in-group. Thus half of the new attachments are likely to be towards an agent from an out-group who is not valued by morally parochial agents. Figure 3-9a shows the average weight on parochial principles such as kin and in-group compared with the broader principles of all people and indirect-reciprocity. We compared the average weight as a function of the number of generations and the proportion of agents generating new attachments (ρ). When $\rho = 0.05$, there is very little cumulative moral change towards indirect-reciprocity and all people. However when $\rho = 0.15$, there is a complete shift towards these broad values

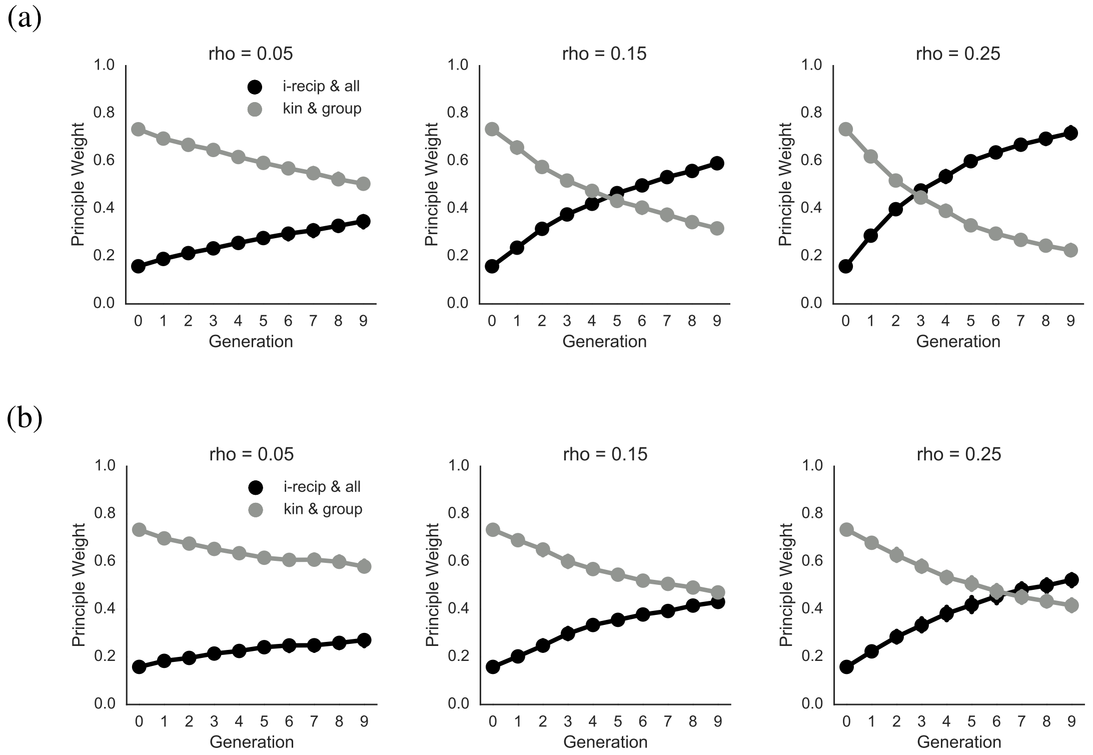


Figure 3-9: Change in the average agent's weighting of parochial vs. impartial moral principles as a function of generation and the proportion of agents (ρ) that develop an attachment (ϕ) for another agent chosen (a) uniformly at random or (b) in proportion to their interaction frequency. The 0th generation is the starting state. As ρ increases, the rate of moral change rapidly increases in (a) but in (b) moral change is significantly inhibited.

but only after many generations. Finally, when $\rho = 0.25$, agents predominantly weigh the impartial principles after only three generations.

In the second set of simulations, agents formed attachments towards other agents proportional to their probability of interacting with that agent. These agents were far less likely to form a new attachment to someone outside of their in-group since they rarely interact and observe the behavior of agents outside of their in-group. Figure 3-9b shows how the moral theories changed under this paradigm. Unlike previous simulations, when $\rho = 0.05$, almost no moral change was observed and after one generation the moral theory remained relatively constant. Even when $\rho = 0.25$ which led to rapid moral change in the previous set of simulations, moral change was slow and the parochial values and impartial values did not cross over until after around ten generations.

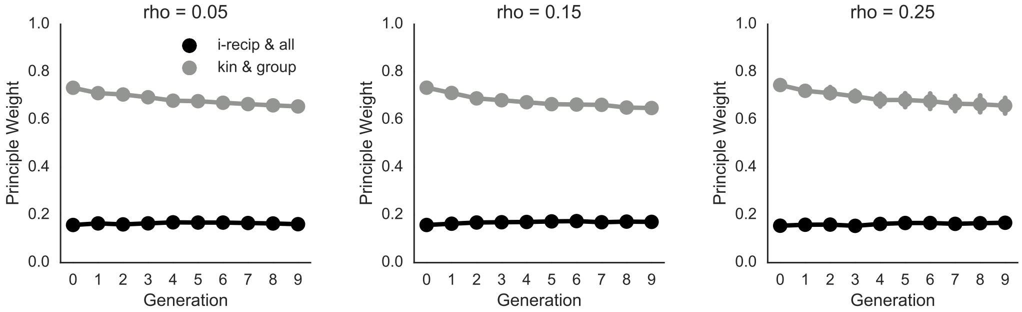


Figure 3-10: Moral change from attachments critically depends on internal alignment. When simulations are run without internal alignment active during learning, there is no significant moral change towards impartial values no matter the proportion of agents (ρ) that develop an attachment for another agent.

To test whether the previous results depended on the internal alignment mechanism, we ran the same simulations as above but without internal alignment active during learning (Figure 3-10). No matter the amount of attachments formed (ρ), there was little to no change in the moral theories demonstrating that moral change based on attachments critically requires internal alignment.

This result could also correspond to being aware of the inconsistency but lacking the meta-value to reduce the conflict, choosing to live with that inconsistency rather than revise one's moral theory (Bennett, 1974). Another possibility is that agents are simply unaware of the inconsistency – people often feel strong attachments for their spouses and neighbors but remain inconsistent. Instead, they must construe the attachments and feelings for their loved ones as incompatible with their moral position. A recent study by Hein, Engelmann, Vollberg, and Tobler (2016) showed that unexpected prosocial behavior from an out-group member elicited a neural signal consistent with a prediction error. These signals could also act as a cue to initiate the process of updating one's moral theory. Furthermore, unequal deserving of moral concern is not always or obviously seen as incompatible with feeling love for specific individuals. Others may be seen as appropriately and rightly occupying different positions in the moral arrangement, and therefore having different rights without necessarily generating any internal alignment. Agents may also be motivated by personal image or other selfish motivations to ignore the inconsistency (Monin, 2007; Monin, Sawyer, & Marquez, 2008).

Can this explain why attitudes about some groups change quickly (e.g., women and homosexuals) but change slowly or not at all for others (e.g., races, religions and nationalities) even once those inconsistencies are pointed out? One possibility is that internal alignment does not operate automatically. Instead, inconsistency may need to be experienced and lived repeatedly to generate moral change through internal alignment. This lack of continued and interactive contact may underlie the cases where moral change is resistant. An intriguing possibility along these lines is the role of literature in spurring moral change (e.g., *Uncle Tom's Cabin*) by activating internal alignment. Literature can humanize a person in morally relevant ways, forcing a reader to experience their inconsistency over and over again. A particularly effective way to generate moral change may be to combine external and internal alignment. A moral exemplar describes and relates their own process of noticing inconsistency and resolving it through internal alignment, simultaneously walking others through their own moral change and encouraging them to do the same.

While we have demonstrated that attachments can in some cases lead to rapid moral change from a parochial moral theory to an impartial one, we now investigate whether attachments selectively generated towards one's in-group towards can change agents that have impartial moral theories into having more parochial moral theories – narrowing the moral circle. Figure 3-11 shows simulations with a society that starts with an impartial moral theory and in each generation agents form attachments with other agents specifically within their in-group. No regression towards parochial values was observed. From these simulations we hypothesize a "moral ratchet effect," since impartial moral theories that value all agents already include valuing those in-group members, no inconsistency arises from those attachments. Thus moral change towards more impartial theories is robust to new positive attachments towards one's in-group and is not expected to lead to moral regression.

The dynamics of these results suggest there may be a critical point for enabling long lasting moral change. When agents were more likely to be exposed to and develop attachments to agents outside of their in-group they quickly revised their moral theories to be consistent with these attachments and developed impartial moral theories. When agents

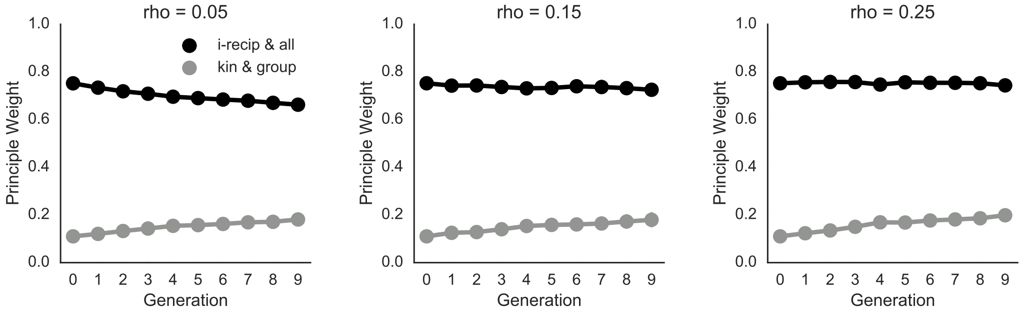


Figure 3-11: Moral change towards impartial values is robust to in-group attachments. Agents started with an impartial moral theory but each generation developed attachments towards others with probability proportional to their interaction frequency. Thus most of these attachment were formed with kin and in-group members. Although attachments were parochial, there was little change in the average moral theory.

were limited in their out-group interaction, their parochial moral theories persisted for far longer. This work suggests that moral learning is a double edged sword: while it is possible to rapidly and reliably acquire a set of abstract principles from limited and sparse data, the values acquired might reflect group biases. Under the right circumstances moral progress can appear rapidly but in other circumstances it fails to cross group boundaries.

3.4 Discussion

We have argued that three principles should be central in a computational framework for understanding moral learning and moral change. First, the commonsense moral knowledge used to make trade-offs between the welfare of different people including oneself can be represented as a recursive utility calculus. This utility calculus weights abstract moral principles and places value on people enabling the evaluation of right and wrong in an infinitude of situations: choosing when to act altruistic or reciprocal, favoring one person or group of people over another, or even making judgments about hypothetical out-of-control trolleys, etc. This abstract representation contrasts with previous formal models of moral learning where the knowledge that supports moral judgment consists of simple behaviors or responses to behavioral reinforcement (M. A. Nowak, 2006; Rand & Nowak, 2013;

Cushman, 2013). Moral knowledge grounded in behaviors rather than abstract principles of valuation cannot generalize flexibly to novel situations.

Second, for moral theories to be culturally learned, learners must be able to infer the moral theories of others, and we showed that hierarchical Bayesian inference provides a powerful mechanism for doing so. Rational inference is needed to figure out which moral principles and reasons drove agents to act in a world where moral behavior and judgments are sparsely observed, noisy and often ambiguous – a "poverty of the stimulus". What a person does in one context gives information about what they will do in other contexts, and learners exploit these regularities to go beyond the data to infer the abstract principles that drive a person to act. The hierarchical Bayesian model exploits regularities in how moral theories are shared between group members to generalize rapidly to new people the agent may have never seen before. In addition to inferring the moral theories of other agents, our model also infers reciprocity relationships which cannot be directly observed. Without the ability to infer abstract theories, learning would be limited to behaviorist models which only care about the observable behavior of others, not their character or reasons for acting.

Finally, having inferred the moral theories of others, learners must choose how to set their own moral theory. We argue that moral learning is guided by meta-values which determine the kinds of moral theories that the learner values holding. Under this model, moral learning is the process of aligning one's moral theories with these meta-values. A meta-value for external alignment, tries to match the learner's moral theory as closely as possible to the inferred moral theories of the people that the learner values. External alignment accounts for the reliability of moral learning from others across generations and gives an account of how agents mix together the moral theories of the many agents they may end up caring about. The richness of this form of cultural learning critically requires both the ability to represent abstract moral theories and infer the moral theories of others. A second meta-value, internal alignment, revises moral theories to make them consistent with attachments and feelings generated from emotional (empathy, love, contact) and deliberative sources (analogies, argumentation, stories) (Allport, 1954; Bloom, 2010; R. Campbell & Kumar, 2012). Our model makes testable predictions about how the different patterns of attachments could affect the dynamics of moral change.

Our core argument is that a full account of moral learning should include at least these three computational principles: moral theories represented in terms of abstract principles grounded in a recursive utility calculus, hierarchical Bayesian inference for rapidly inferring the moral theories of others, and learning by value alignment both externally to the values of others and internally through reducing inconsistency. Our main results take the form of a series of simulations based on a particular implementation of these principles, but we stress that our specific implementation is unlikely to be fully correct and is certainty not complete. Many of the specific quantitative modeling choices we made (for instance, the choice of squared-error as opposed to absolute difference for the learner's cost function on weights, or the choice of a normal distribution as the prior over weights) do not affect the main results and we are not committed to them specifically. Instead, we want to argue for and explain the value of several computational principles more broadly in moral learning, and we hope that their instantiation in a specific computational model can complement more qualitative accounts of moral learning and moral change (Singer, 1981; D. A. Pizarro, Detweiler-Bedell, & Bloom, 2006; Pinker, 2011; Mikhail, 2011). Ultimately, we hope that understanding the mechanisms of moral change at this level can ultimately be valuable in implementing the changes we would like to see in our societies – or in understanding when moral progress is likely to be slower than we would like.

Given that this is a first attempt at using these quantitative tools in the moral domain there are still many possible extensions we hope to address in future work. In this work learners received data in the form of moral judgments and behaviors, however external alignment is sufficiently general to learn from other types of data such as explicit declarations of values. For example, a value statement such as "Family comes first!" could be encoded as a qualitative constraint on the ordering of weights for different moral principles, i.e., the weight on `kin` should be higher than on other principles. It can also be used to learn from punishment and praise. Consider the difference about what is learned when punished by an anonymous person versus someone you love. In part, the decision to punish gives information about the punisher's own moral theory. If the punisher is someone who the learner cares about it can lead to moral updating through external alignment rather than behavioral reinforcement.

Other extensions could integrate our model with recent work which has shown how deontological principles (of the form "do not X" or "do not intend X" regardless of the consequences) could be learned (Nichols et al., 2016; Ayars & Nichols, 2017) or emerge from choice algorithms (Crockett, 2013; Cushman, 2013). Learners are also expected to learn how different "base" moral goods and evils contribute to the welfare of individuals or even what counts as moral. Differences in what counts as moral is already known to vary across cultures and individuals (Graham et al., 2009, 2016). In our model this would correspond to learning the form and weight of different components in the $R(s)$ function. In this work we treated all moral goods as having a shared currency ("utility") but people may act as if there are multiple sets of value, different currencies that cannot be directly interchanged (Baron & Spranca, 1997; Baron & Leshner, 2000; Tetlock, Kristel, Elson, Green, & Lerner, 2000). Finally, these source of moral value may also compete with mundane and non-moral values (Tetlock, 2003). We leave these challenges for future work.

Much more can also be said about the structure of moral principles in our framework. Group membership is often combinatorially complex where each agent may be a member of multiple groups some observable and others not. Some groups are defined top-down by external factors such as race, religion, gender, or location while others are defined bottom-up such as based on a similarity of values (moral and non-moral). While in this work, we showed how the priors on the values of group members can speed up the inference of the values of individuals, it can also speed up an inference of who is in what group by exploiting knowledge of their values. Groups are themselves dynamic and future work should integrate models of group formation with the dynamics of moral theory learning (Gray et al., 2014).

Furthermore, in the simulations we studied, there were only two groups which were of equal size and which shared similar values. We could ask, for example, whether a learner with a caregiver who holds a minority moral theory is as likely to spread that theory as one with a caregiver who holds a theory held by the majority? When are minority values likely to be assimilated into the majority after a few generations, and when do they become stable? Or consider the effects of ambiguous moral inference on moral change. A person in one group may show a few cooperative interactions with members of another group, which could reflect a low in-group bias and high impartiality. But these actions could also come about from a high in-group bias together with some specific valuation of a small number of out-group members, either through highly weighted direct reciprocity links or intuitive feelings. Others may not know how to interpret their actions, and indeed the individual may themselves be confused or self-deceptive, as exemplified by the classic excuse, "I'm not racist! Some of my best friends are black!". How might these ambiguities speed or slow the rate of change towards impartial indirect-reciprocity in the expanding-circle scenarios we discussed above?

While in this work we mainly explored how the moral principles are abstract with respect to individuals and groups, we observe that such principles are also abstract to situational context (Fiske, 1992). In some contexts one might be justified in acting mostly in one's own interests or the interest of one's loved ones while in another context selfless behavior may be obligated. For example, it may be acceptable to give higher weight to one's own child under most circumstances, but when acting as a school chaperone this duty is extended equally to all the children. Furthermore, there are exchanges of welfare based on merit, effort or punishment which require a notion of proportionality that our representation does not capture (Rai & Fiske, 2011).

We hope in future work to be able to say more about where these moral principles cognitively originate. Some have argued that children might have an innate understanding of even the more sophisticated reciprocity based moral principles (Hamlin, 2013). Another possibility is that these principles come from an even more abstract generative model of moral and social behavior, either embedded in the roots of societies through something like an "initial position" bargain (Rawls, 1971; Binmore, 1998) or implemented in a more online fashion in individuals' "virtual bargaining" with each other (De Cote & Littman, 2008; Misyak, Melkonyan, Zeitoun, & Chater, 2014; Kleiman-Weiner et al., 2016). Evolutionary mechanisms (cultural or biological) which favored groups that followed these principles, because of how they promote cooperation and the advantage cooperation bestows to groups and their members, are also likely contributors (Rand & Nowak, 2013; Greene, 2014). Our work here is complementary to all these proposals, and we would like to explore further how it could integrate with each of them.

Finally, if we are going to build artificial agents that can act with us, act on our behalf and make sense of our actions, they will need to understand our moral values (Wiener, 1960; Bostrom, 2014). Our model suggests one route for achieving that understanding: We could build machines that learn values as we propose humans do, by starting with a broad set of abstract moral principles and learning to weight those principles based on meta-values which depend in part on the values of the humans that the machine interacts with or observes. This proposal fits well with mechanisms of value alignment via cooperative inverse reinforcement learning (Hadfield-Menell, Russell, Abbeel, & Dragan, 2016) that have been proposed for building beneficial, human-centric AI systems. We can choose how much of morality should be built in to these machines and how much should be learned from observation and experience. With too little abstraction built in (such as trying to learn the α directly), the machine will learn too slowly and will not robustly generalize to new people and situations. With too much structure and constraints, the restricted theory may be unable to capture the diversity and richness of the true moral theories used by people. The model presented here is just one point on this spectrum which trades off complexity and learnability. The prospect of building machines that learn morality from people hints at the possibility of "active" moral learning. Can a learner, child or machine ask questions about ambiguous cases (perhaps similar to those pondered by philosophers) to speed up the process of moral learning?

In conclusion, learning a commonsense moral theory, like learning a language, turns out to require a surprisingly sophisticated computational toolkit. This is true if we seek to understand how moral knowledge is acquired, particularly the type of moral knowledge that generalizes flexibly to an unbounded range of situations, and that involves interactions with others we barely know or have never met. Understanding moral learning in computational terms illuminates the cognitive richness of our moral minds, and helps us to understand how our societies might have come to the moral values we hold – and where we might be going.

3.5 Simulation Details

In this work we consider two types of decision contexts: one where the actor traded off her own welfare for that of another person, and one where the actor traded off the welfare of one agent for the welfare of another. For the first type of decision context, an actor chose between an allocation of welfare of 0 to herself and 0 to the other agent or an allocation of $-A$ to herself and $A+B$ to the other agent where A and B were independently resampled from an exponential distribution with mean 10 for each decision. Thus in these decisions an agent chooses between doing nothing, or paying a cost ($-A$) to give a larger benefit to another agent ($A+B$). The larger the ratio of the samples (B/A) the greater the joint utility of choosing the prosocial option.

For the second type of decision context, the actor chose between A welfare for one agent and $A+B$ welfare for another agent with no impact on the actors own welfare. In this context, the actor is choosing which person should be given the allocation and the agent not chosen gets nothing. A was resampled from an exponential distribution with mean 10 and B was independently sampled from the same distribution as A with probability 0.5 and set to 0 with probability 0.5. Although there are only two decision contexts, since the actual welfare trade off is newly sampled for each choice, no decision is exactly like any other.

To generate observations for learning, we first sampled an actor and affected agents from the previous generation of agents and a decision context with values for A and B. Then a choice or judgment was generated by sampling from the distribution shown in equation (3.3) with $\beta = 5$. Each learner observed a unique set of decisions and judgments from different actors. We assumed that the observed agents have already reached an equilibrium in learning i.e., the agents which generate observations are not themselves learning. Due to this assumption each observation of a decision is independent.

Maximum a posteriori probability (MAP) inference for the conditional on the observations ($P(W|(a_i^0, s^0), \ldots, (a_i^T, s^T))$) was estimated using an EM-like inference algorithm that iterated between optimizing the weights W_i of each agent i, the group average weightings $W^g_{\texttt{norm}}$, and samples from the two reciprocity relationships ($P(f^{\texttt{d-recip}}, f^{\texttt{i-recip}}|H)$). In all simulations we used $\lambda = 1$ for $P(f^{\texttt{d-recip}})$, $p = 0.5$ for $P(f^{i-recip})$ and $\Sigma^g = \mathbf{I}$ for all g.

3.6 Extending the Utility Calculus

Here we explore possible extensions to the representations of moral theories which demonstrate the richness of the utility calculus. While we considered recursive utility calculus where prosocial moral theories the level-1 theory is composed from self-valuing level-0 moral theories. We can iteratively apply recursive valuation to generate utility functions that allow for higher-order preferences. The level-k utility function is:

$$U_i^k(s) = (1-\gamma_i^k)U_i^{k-1}(s) + \gamma_i^k \sum_{\substack{j \in N \\ j \neq i}} \alpha_{i,j} U_j^{k-1}(s)$$

An agent with a level-k moral theory goes beyond just valuing people but also includes recursively valuing the people they value and so on. If γ_i^k decreases as a function of k (i.e., $\gamma_i^k < \gamma_i^{k-1}$), higher orders of recursive valuation become progressively less important.

We can also consider a moral theory that is not just dependent on the expected state and outcome but also dependent on properties of the action itself. We can abstractly include these prohibitions by modifying the base utility function.

$$U_i^0(s,a) = R_i(s) - \delta_i D_i(a)$$

where $D(a)$ is a function that returns the degree to which an action violates a deontological rule that agent i cares about. Since intentions can be inferred from actions (Mikhail, 2007; Kleiman-Weiner et al., 2015), these constraints could include restrictions on intention such as the doctrine of double effect or other specific forbidden actions (Tetlock et al., 2000; Haidt, 2007). Importantly, these norms are limited to those that only depend on the action (and what can be inferred from the action), without reference to the consequence. These deontological norms are integrated with the rest of the moral theory with δ_i controlling the relative degree that agent i takes into account deontological rules compared to outcomes (Kleiman-Weiner et al., 2015; Nichols & Mallon, 2006). Recent research has made progress on learning this function from experience (Cushman, 2013; Nichols et al., 2016; Ayars & Nichols, 2017). Once this new base utility function (U^0) enters the level-k

recursion, if agent i values the utility of agent j through $\alpha_{i,j}$, than i will also care about the deontological prohibitions that agent j cares about. To use these utility functions which depend on actions as well as states requires simply substituting $U(s')$ in equation (3.2) for $U(s',a)$.

Chapter 4

Learning to Cooperate and Compete

4.1 Introduction

Our most important relationships involve understanding when to cooperate and when to compete. From siblings to coworkers, humans rely on both planning and context to know which situations they should cooperate in and which they should compete in (Galinsky & Schweitzer, 2015; Rand & Nowak, 2013). And yet in real life, unlike a behavior economics experiment, cooperation and competition are abstract with respect to a given situation. A cooperative or competitive interaction unfolds over time – there isn't a single moment where competition or cooperation "happens". Even if the decision to cooperate or compete has been made, efficiently implementing those strategies can be difficult. A person determined to cooperate and knowing what the other person wants will have to develop a

Matrix-Form Games

		Yellow	
		Cooperate	Compete
Blue	Cooperate	7,7	-1,8
	Compete	8, -1	4,4

Figure 4-1: A social dilemma written as a normal-form game. The numbers in each square specify the payoff in terms of utility to the blue and yellow player respectively for choosing the action corresponding to that square's row and column. If both agents choose cooperate they will collectively be better well off than if they both choose compete. However in any single interaction, either agent would be materially better off by choosing to compete.

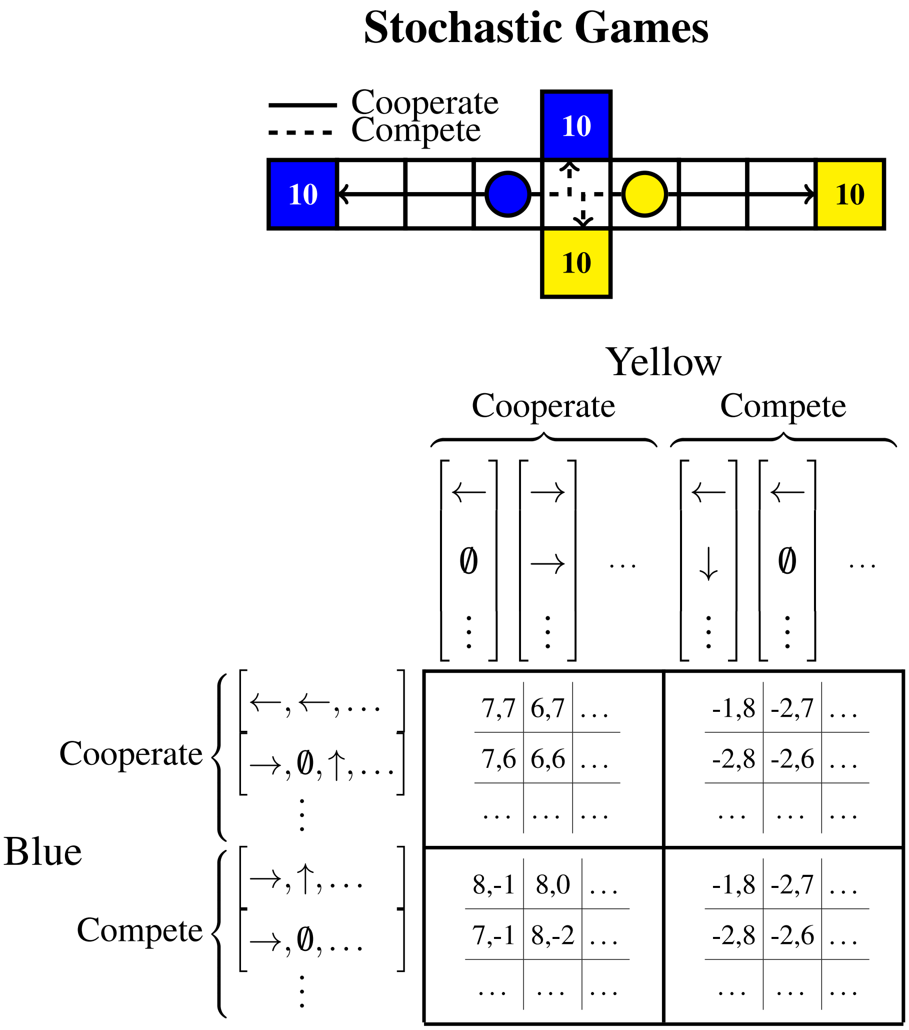


Figure 4-2: Two-player stochastic games. (top) Grid form representation of the stochastic game. The arrows show example strategies that can be used to realize both cooperative and competitive outcomes. (bottom) Matrix representation of the strategy space, with low-level strategies sorted by a high-level goal. The arrows correspond to moving in a specific direction and the ∅ corresponds to waiting. Note that the action space is effectively unbounded but the strategies naturally cluster into a small number of high-level goals. If both agents go to the sides then they will both score the reward but if they fight for the middle in hopes of using less moves they will collide and only one will get any reward.

detailed plan of action to realize that cooperative intention. Likewise for a person intent on competing. In this work we aim to bridge high-level strategic decision making over abstract social goals such as cooperation and competition with low-level planning over actions to actually realize those goals.

The ability to form these hierarchical joint intentions is a key component of social behavior. The motivated instinct to both infer and evaluate complex social plans emerges in early childhood (Warneken & Tomasello, 2006; Hamann, Warneken, Greenberg, & Tomasello, 2011). Young children not only rapidly infer the goals of other agents, but spontaneously execute complex plans to cooperate with others. For instance, a cooperative intention might generalize to include not just the low-level details of a joint task but also tell how to share the spoils. The ability to infer the intentions of others and participate in a dynamic joint endeavor (sometimes called the "we-mode") is thought to be a key building block of large scale collaborative culture (Tomasello et al., 2005).

4.1.1 Naturalistic Games

Game-theoretic investigations of social behavior often represent strategic interactions as *matrix-form* games like the one shown in Figure 4-1. In these games, the rows and columns correspond to the action space of the two players and the cells describe the payoffs to each agent that would result from those actions. While useful as a succinct representation of a social decision, these games lack the ecological validity of real social decisions which require planning across space and time. When presented to participants, it can be difficult to extract the right information and even after significant training, many people don't even look at the payoffs most relevant for strategic reasoning (Costa-Gomes, Crawford, & Broseta, 2001). When the number of decisions grows beyond two decisions per agent, these problems are exacerbated.

Instead we use a paradigm commonly deployed in multi-agent systems research which has not been explored behaviorally (De Cote & Littman, 2008). In this paradigm, strategic interactions are represented as naturalistic spatial environments that people play intuitively like video-games. Figure 4-2 shows an example of one of these multi-agent planning environments that is conceptually related to the social dilemma shown in Figure 4-1. Unlike the matrix-form game, these environments also require low-level planning over spatial actions to realize a strategic goal. The action space of these games is much larger than those typically studied in matrix-form games but the strategies are still intuitive.

Each player controls the movement of one of the colored circles. On each turn players choose to either move their circle into an adjacent square (not including diagonal moves) or to remain in the same position. Attempting to move is costly resulted in the loss of one point. Choosing to remain in the same position did not incur any cost. Both players select an action during the same turn and their positions are updated simultaneously. Each square can only be occupied by one player at a time so if both players try to move to the same square, one of the players chosen by chance will enter the contested square while the other remains in place. However both pay the cost for attempting to move. If one player stays in the same position and the other player tries to move into their square, no movement occurs. Finally, players cannot move through each other and switch places.

The colored squares are the goals. When either player reaches a square with the same color as their avatar, that player receives ten points and the round ends. Thus the only way for both players to receive points is if they both enter squares that match their avatar's color on the same turn. These dynamics were chosen to be identical to those in (De Cote & Littman, 2008) so that our data can also be compared to the models of that work. Because each interaction generates data about both the action plan and the payouts, we can use these games to start to investigate the mechanisms people use to coordinate on cooperative and competitive outcomes. Furthermore, they allow us to study how humans innovate to find these strategies out of such a large possible space of action plans.

4.2 Model

4.2.1 Hierarchical Social Planning

We develop a hierarchical model of strategic planning that unifies low-level action planning with high-level strategic reasoning and allows for learning across both levels. In brief, agents have two "modes" of low-level planning: a cooperative mode and a competitive mode. These two modes are connected through a high-level strategic planner that determines which mode should be deployed based on previous interactions. After each round, agents use Bayesian theory-of-mind to determine whether or not the other agent's low-level actions are consistent with the cooperative planning mode vs. the competitive planning mode. The agent can then condition their own next actions on the inferred high-level intentions of the other agent realizing a sophisticated strategic response.

Both modes include forms of model-based learning which allows for learning to generalize across environments as well as model-free reinforcement of actions. In this work we focus specifically on the high-level goals of cooperation and competition but other high-level goals such as teaching, punishing or communication are also relevant in these games and will be investigated in future work. The challenge of hierarchical planning is to link these high-level goals to a lower-level plan of action.

Our work builds on and is inspired by classical formalisms of intention and joint planning from the AI literature (Levesque, Cohen, & Nunes, 1990; Grosz & Kraus, 1996) as well as more modern formulations for planning under uncertainty such as DEC-POMDPs and I-POMDPs (Gmytrasiewicz & Doshi, 2005; Gal & Pfeffer, 2008; De Cote & Littman, 2008). However the earlier models do not handle uncertainty in a probabilistic way and hence struggle with quantitative predictions about behavior while the later are often intractable over long planning horizons and don't explicitly represent abstract social goals.

We briefly introduce stochastic games following the notation of De Cote and Littman (2008) and then discuss repeated stochastic games. A two-player stochastic game is: $\langle S, s_0, A_1, A_2, T, U_1, U_2, \gamma \rangle$ where S is the set of all possible states with $s_0 \in S$ the starting state. Each agent can choose from a set of actions A_1 and A_2 which together form a joint action space $A_1 \times A_2$. The state-transition function, $T(s, a_1, a_2) = P(s'|s, a_1, a_2)$ maps a state and joint action to a distribution over new states. The utility functions of the two agents $U(s', s, a_1, a_2) = R$ describe the agent's goals in terms of quantitative costs and rewards. Finally $0 \leq \gamma_{\texttt{game}} \leq 1$ is the discount rate of reward. In repeated stochastic games, a series of stochastic games are played one after another in succession between the same pair of players. We now discuss the cooperative and competitive modes of planning in detail.

4.2.2 Cooperative Planning

Since there is no specific action that corresponds to cooperation in these stochastic games (all actions are spatial movements), we develop an abstract notion of cooperation which generalizes across contexts. We postulate that a cooperative action is one that is good for the group i.e., efficiently maximizes the utility of all agents. Since under this assumption, the goal of cooperation is to rationally achieve a group goal, we consider a *group-agent* that optimizes a utility function composed of the utility of all agents (Sugden, 1993, 2003; De Cote & Littman, 2008).

Computationally, we represent this group utility function as a linear weighting of the utility of the two agents: $U^G = (w)U_1 + (1-w)U_2$ where $w \in [0,1]$ controls how the two agents are relatively valued by the group-agent. For example when $w = 0.5$ the group-agent impartially weighs the utility of both agents equally. We are not implying that this group-agent actually exists but rather that each player can simulate the same group-agent by taking an objective view of the planning environment outside and separate of their own personal goals (Nagel, 1986). Since the "view from nowhere" is a common view for these agents, it avoids the outguessing regress of thinking about others, and is instead a space where the details of planning can be hashed out for coordination using what is commonly known all agents. We note that this utility function can include other social preference such as inequality aversion or merit based allocations.

Since the group-agent can directly control the actions of both players (like a "we" agent), it can treat the stochastic game as a single-agent MDP. Rational planning over joint actions (a_1, a_2) is achieved through value-iteration:

$$
\begin{aligned}
P(a_1,a_2|s) &= \pi^G(s,a_1,a_2) \propto e^{(\beta Q^G(s,a_1,a_2))} \\
Q^G(s,a_1,a_2) &= \sum_{s'} P(s'|s,a_1,a_2)[U^G(s',s,a_1,a_2) + \\
&\qquad\qquad \gamma \max_{(a'_1,a'_2)} Q^G(s',a'_1,a'_2)]
\end{aligned}
$$

where the group-agent policy, $\pi^G(s)$, is to choose actions with probability proportional to their future expected utility. A high value of β means that the group-agent is more likely to select the action with the highest Q-value and a low value of β means that the group-agent is more likely to select suboptimal-actions. In all experiments we used a relatively high value of $\beta = 4$. We note that π^G is not only a policy for action, but also includes the future-oriented intentions of what the two agents *should* do once they get to a new state. These intentions include how to recover from failed coordination attempts. We used a discount rate of $\gamma = 0.9$ in all the models presented here.

Although each agent might consider the policy of the group-agent, the individual agents can only control their own actions. To transform this group-agent policy into an individual policy, individual agents marginalize out the actions of the other player from the joint policy: $\pi_1^G(s, a_1) = \sum_{a_2} \pi^G(s, a_1, a_2)$ and $\pi_2^G(s, a_2) = \sum_{a_1} \pi^G(s, a_1, a_2)$. These policies contain intertwined intentions, not only an *intention to* take a specific action but also the *intention that* the other agent reach certain states. This "meshing" of plans between the two agents has been called a key component of joint and shared intentionality (M. E. Bratman, 1993, 2014). Unlike social preference based accounts of cooperative behavior where each agent individually plans to maximize joint utility, in this account, cooperation is a built in cognitive feature of planning itself – agents *plan together*.

When there is a single unambiguous action for both players that maximizes joint utility, coordination is readily achieved. However in the environments we investigate, there are often multiple actions that can generate optimal rewards for the group-agent. We now discuss two mechanisms for learning social norms that can break these symmetries and lead to robust coordination on a single jointly optimal plan.

We first consider the case where two different actions are equally good from the perspective of a group-agent that weighs the utility of the two agents equally but the rewards will be allocated unequally. For example, consider game (C) in Figure 4-3 where one agent needs to go around the other. Because moving costs 1 point, the agent who goes around the other will only earn 7 points while the agent who waits will earn 9 points. From the perspective of the group-agent with $w = 0.5$, it doesn't matter who goes around since the joint utility is equal. However if one agent was favored over the other ($w \neq 0.5$) this symmetry would be broken and the disfavored agent would take the long route. Thus prior knowledge about asymmetries in how the group should operate can lead to more robust coordination although potentially at the cost of less fair cooperation.

The two agents may start with a different prior on the value of w and thus when simulating the group-agent will fail to coordinate. Consider the case where both agents think they should be valued more than the other and hence expect the other player to go around them. We propose a mechanism based on “virtual bargaining” accounts of social choice that lead to each agent’s w to converge over time to the same value without any explicit communication (Binmore, 1998; Misyak et al., 2014). After each interaction, agents can infer the w that best explains the joint behavior of their previous interaction: $P(w|H) \propto P(H|w)P(w)$ where H are the data from previous interactions and the likelihood of those interactions is defined by the marginalized joint policies generated from planning with a specific w: π_1^G and π_2^G. In our analysis, each agent starts out with a prior of $w = 0.5$ and updates it after each round based on the inferred w of the previous interaction. Thus over time w will converge and as predicted by the theory of virtual bargaining, more patient agents who insist on the advantage will gain a greater share of the joint reward in future coordinated interactions where an equitable split isn’t possible. For example, if in a previous interaction agent 1 took a more costly route, then in the next round agent 1 will be more likely to take the costly route again generating a social norm for cooperative coordination. Since w is an input to the planning process itself, it allows for generalizing these norms to new environments.

Finally, in some environments, there are multiple plans that are equally good for both agents, creating a different type of symmetry which cannot be broken by w. For example, the decision to go clockwise or counterclockwise in game (A) of Figure 4-3 is equally good for both players as long as they both go in the same direction. To capture the intuition that once agents successfully coordinate, they should continue to coordinate in that way e.g., after luckily choosing to go clockwise in game (A), they will go clockwise again on the next round, agents learn a function $N^G(s, a_1, a_2)$ based on the frequency of previous joint actions which is added to the state-action Q^G-value used by the group-agent. This norm based reinforcement affects the policies of the individual agents through marginalization. The norms reinforced by this mechanism do not generalize across environments although feature based norms can generalize when there are features in common between two environments e.g., see Ho et al. in this years proceedings.

4.2.3 Competitive Planning

As before, in these stochastic games there is no action that directly corresponds to "compete". Instead, we ground competitive planning as each agent attempting to maximize their individual utility under the assumption that the other agent is doing the same. To tractably realize this game-theoretic best-response, we extend the cognitive hierarchy / level-K formalism used in behavioral game theory to temporally extended polices instead of just actions (C. F. Camerer, Ho, & Chong, 2004). In brief, a level-K agent best responds to a level-$(K-1)$ agent which grounds out in the level-0 agent. Specification of the level-0 agent is sufficient to specify the full hierarchy.

In this work we use a level-0 agent that doesn't consider the existence of the other player and tries to efficiently reach her goal without taking any strategic consideration of how the other player might affect her progress. This level-0 agent is more naturalistic than randomly acting agents which are commonly used in behavioral modeling (C. F. Camerer et al., 2004; Yoshida et al., 2008). A level-0 agent of this type only makes sense in these naturalistic environments since one can easily imagine acting alone unlike in matrix-form games. The level-0 agent for player i is:

$$P(a_i|s,k=0) = \pi_i^0(s) \propto e^{\beta Q_i^0(s,a_i)}$$

$$Q_i^0(s,a_i) = \sum_{s'} P(s'|s,a_i)(U_i(s,a_i,s') + \gamma \max_{a_i'} Q_i^0(s',a_i'))$$

where $P(s'|s,a_i)$ represents transition dynamics that do not depend on the other player. Having defined the level-0 player we can recursively define all of the other levels in the hierarchy in terms of lower levels:

$$P(a_i|s,k) = \pi_i^k(s) \propto e^{\beta Q_i^k(s,a_i)}$$
$$Q_i^k(s,a_i) = \sum_{s'} P(s'|s,a_i)(U(s,a_i,s') + \gamma \max_{a_i'} Q_i^k(s',a_i')))$$

Since the other agent is treated as a knowable stochastic part of the environment, the dynamics of the other player are encapsulated in $P(s'|s,a_i)$ which are marginalized out using the $k-1$ player: $P(s'|s,a_i) = \sum_{a_{-i}} P(s'|s,a_i,a_{-i})P(a_{-i}|s,k=k-1)$ where $-i$ is a shorthand to refer to the "other" player. Because of the maximization operator, a level-K agent implements a best response to a level-$K-1$ agent. Thus zeroth-order agents have their own goals but ignore the other player, first-order agents act on their own goals but assume that the other agent is ignoring their existence and so on. In our experiments we used $K=1$ although results were similar with higher values of K.

Even when competitively planning, agents can still improve their behavior through learning and can even develop certain conventions when they serve mutual self-interest such as symmetry breaking in coordination games. Again we consider two mechanisms. The first mechanism improves agent i's model of agent $-i$ by using the frequency of i's previously successful behavior to modify the state-action Q-values of $-i$ such that previously successful action are more likely to occur again. This model-based mechanism, improves agent i's policy since she will best-respond to a more accurate model of agent $-i$. The second mechanism is model-free reinforcement of player i's state-action Q-values when player i herself successfully reaches a goal. Neither of these norms trivially generalize across different planning environments that don't share states.

4.2.4 Coordinating Cooperation and Competition

Finally, we describe how agents can use both the cooperative and competitive modes of planning to decide whether to cooperate or compete. Since these modes of planning abstract away the details of cooperation and competition, high-level strategic planning can use these low-level planners without considering their details. Agents first use these planning modes to infer the high-level intention I of the other player (i.e., their planning mode) using Bayesian theory-of-mind: $P(I|D) \propto P(D|I)P(I)$ where $P(D|I)$ are just the cooperative or competitive policies. This probabilistic approach is justified because intentions can be ambiguous. For instance, when both agents reach the goal in a coordination game it could just be because of luck so the behavior isn't very diagnostic of the intention. Yet in social dilemma only the cooperative intention is consistent with behavior where both reach the goal. Using these inferred strategic intentions, a high-level planner can take a simple and intuitive form such as reciprocal cooperation (e.g., tit-for-tat) or reinforcement learning at the level of strategy rather than actions (Fudenberg & Levine, 1998).

4.3 Behavioral Experiments

We developed client/server software that allows for real-time interactions between two participants randomly matched through mTurk. All participants went through a short single player tutorial that familiarized them with the controls of the games, the dynamics of the game environment, the costs of movement and value of the goals. After the tutorial, pairs of participants were matched together and played 30 rounds of the same game with the same partner. Subjects were not told the exact number of rounds they would play together in order to prevent horizon effects from backward induction. Once both participants submitted moves, the game state and score were updated and the process continued until the end of the round. Participants had 30 second for each move and the game ended if a participant exceeded their 30 second time bank two moves in a row. We only analyzed data from complete interactions where the pair of participants completed all 30 rounds of the game together. All experiments were incentivized with bonuses proportional to the number of points accumulated.

To compare model predictions with human behavior, we first focused on analyzing whether or not both players reached a goal on a given round, a behavioral signature of cooperation in these games. For each pair of participants, the model observes the interaction in the previous rounds, performs inference on the latent high-level goal and social norms, and samples a prediction for the behavior of the pair in the next round. We compare this sampled prediction with actual human behavior to assess model performance. The same

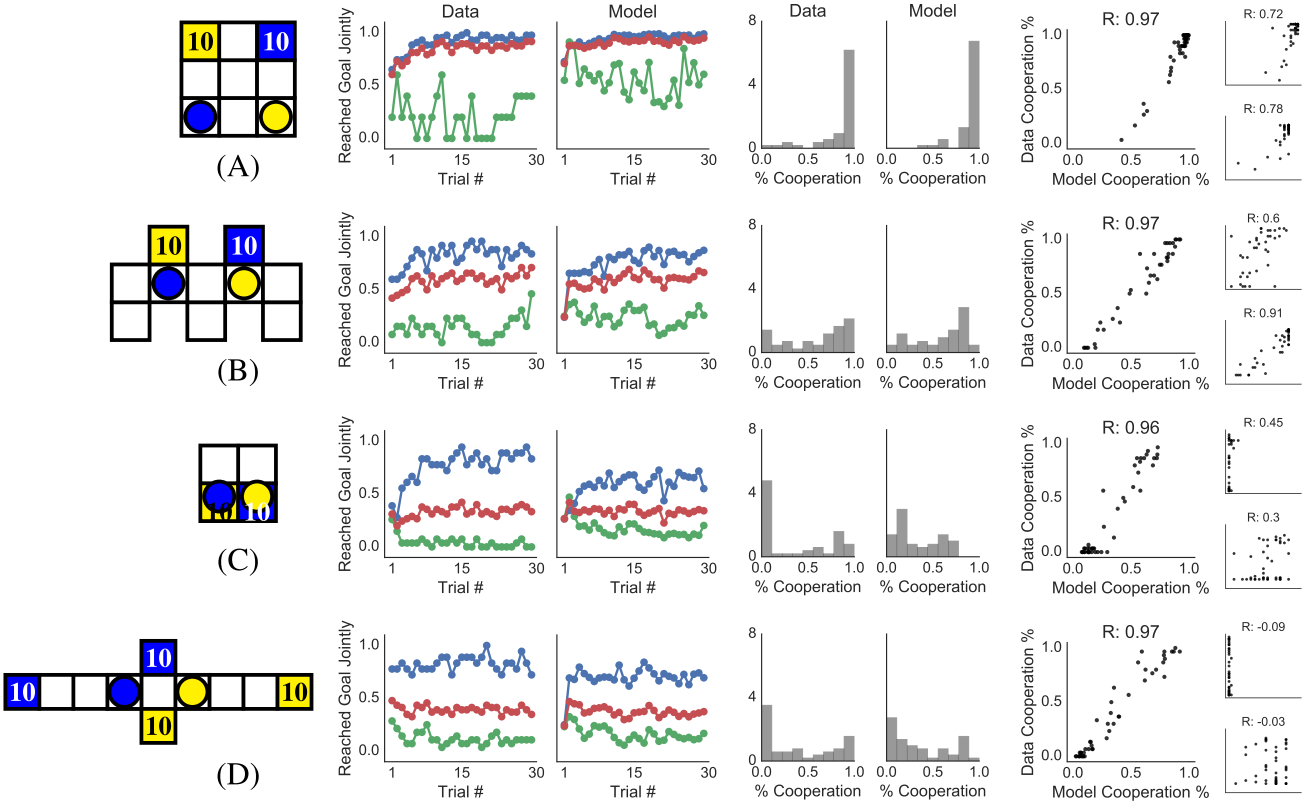


Figure 4-3: Participant data and model predictions for four environments. Each row shows data and model predictions for the environment in column 1 which was repeated 30 times. Rows 1 and 2 are coordination games and rows 3 and 4 are social dilemmas. Column 2 shows the average rate of cooperation for each round of play averaged over the high-cooperating cluster of participants (blue), low-cooperating cluster of participants (green) and all participants (red). Column 3 are histograms of the proportion of cooperation for all pairs of participants. Column 4 quantifies the model predictions where each point represents the frequency of cooperation for a given dyad observed in the data and as predicted by the model. The inset shows correlations of the two lesioned models with the same human data: (top) only compete (bottom) only cooperate.

model parameters were used for all pairs of participants.

Figure 4-3 shows the results of the behavioral experiments and the model predictions for four environments (≈ 50 participant pairs per environment), two coordination games and two social dilemma. Since model predictions were made at the level of each pair of participants, averaging the behavior and model predictions across dyads obscures individual differences in the dynamics of cooperative and competitive learning. To investigate the model predictions in a more fine-grained way, we used unsupervised clustering to split the pairs of participants into two group. In short, for each pair of participants we construct a 30-dimensional binary vector where each dimension corresponds to one of the 30 rounds. Each element is set to one if both participants reached a goal in the round corresponding to that dimension and set to zero otherwise. We ran K-means clustering with $K = 2$ which split the data into a high-cooperating cluster and a low-cooperating cluster allowing for better visualization of the data and model prediction and gave some rough indication about the model ability to handle individual differences.

In all four environments, some of the pairs converged on a cooperative plan but the incentive structure of the game i.e., whether or not the game was a coordination game or social dilemma affected the likelihood that both participants jointly reached a goal. Overall, participants jointly reached the goal more frequently in coordination games than in the social dilemma. As shown in Figure 4-3 the model qualitatively captures the rate of cooperation and competition in both the high-cooperating cluster and the low-cooperating cluster as well as the average over all participants. Another coarse measure of behavior in these games is the distribution of the frequency of cooperative behavior across pairs of participants. In coordination games, the distribution was left-skewed and in social dilemma the distribution was right-skewed. These distributions were captured both qualitatively and quantitatively across these games by the model.

We compared the full model which included both modes of planning and strategic reasoning over those two modes with two lesioned models which just used one of the two planning modes. One lesioned model always used the competitive planning mode and the other lesioned model always used the cooperative planning mode. Overall, neither lesioned model could capture the rates of cooperation between the two clusters and qualita-

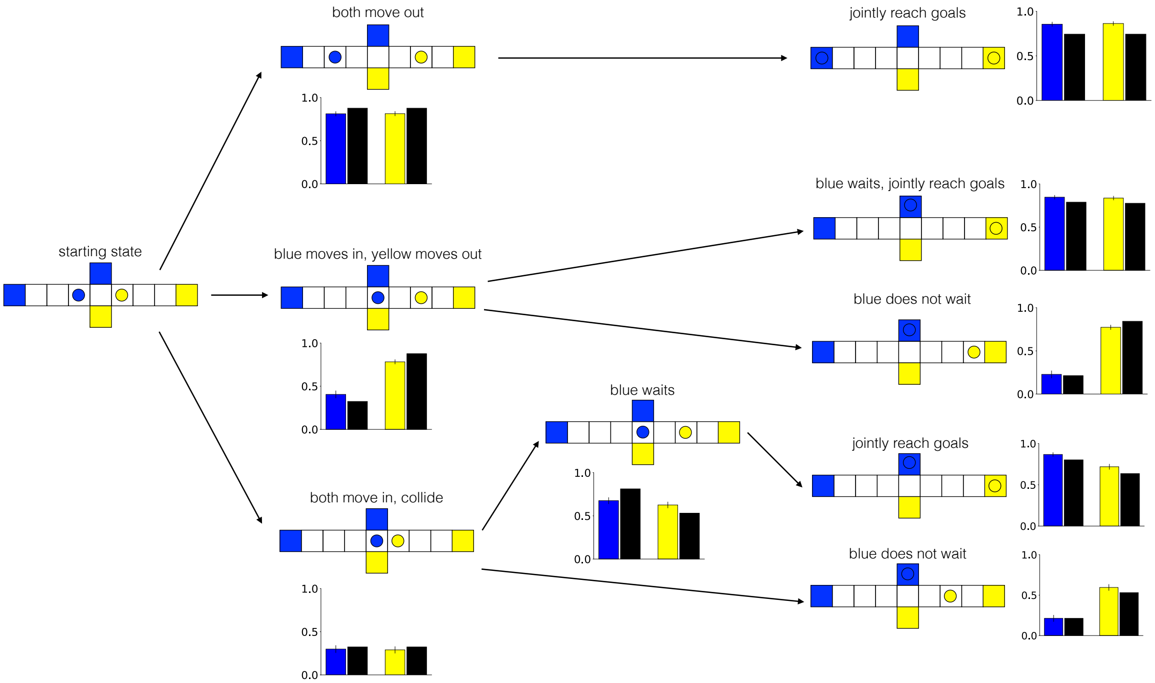


Figure 4-4: Attribution of cooperative intentions in a social dilemma. The blue and yellow bars show the attribution of cooperativeness for blue and yellow agents respectively and the black bars show the model predictions. For all bars, 1 is definitely cooperative and 0 is definitely not cooperative. Error bars show the standard errors of the mean.

tively failed to explain the distribution of cooperative behavior in each game. Both lesioned models failed to predict the dynamics of strategic reasoning between cooperation and competition in social dilemma and had weaker correlation with participants' behavior in the coordination games.

4.3.1 Friend or Foe Inference

Finally, we investigated directly whether or not the abstract planning programs studied here can also act as models of social intention attribution. In particular we asked whether or not people reliably attribution a cooperative intention from just a few observations of behavior even when their behavior was ambiguous. Figure 4-4 shows these inferences in a prisoners dilemma like stochastic game. On the top row both players move towards the outer goals. Just this single act alone was sufficient to infer a joint intention to cooperate. The inference was only reinforced when both agents reached the goals simultaneously. In the middle row, blue moves towards the middle which is interpreted as slightly competitive.

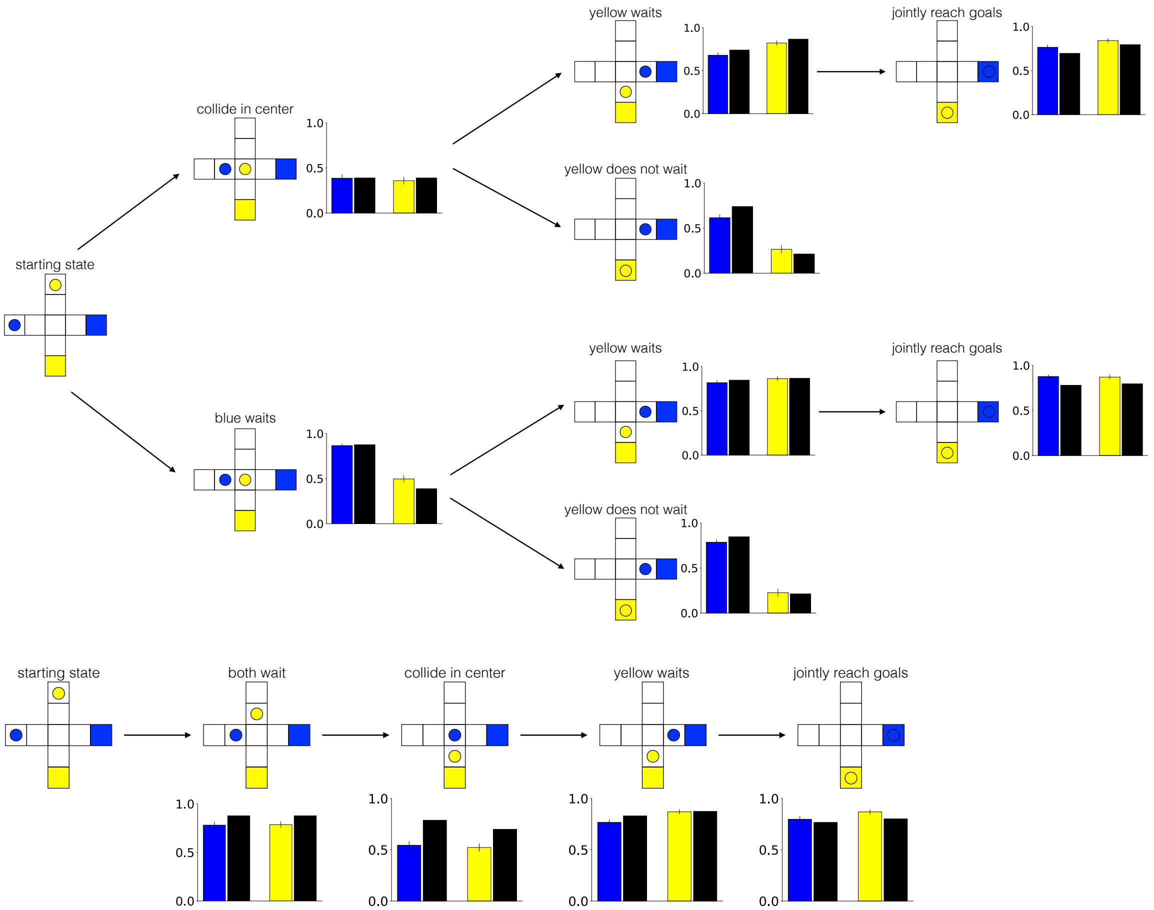


Figure 4-5: Attribution of cooperative intentions in a social dilemma requiring coordination inspired by a four way stop. The blue and yellow bars show the attribution of cooperativeness for blue and yellow agents respectively and the black bars show the model predictions. For all bars, 1 is definitely cooperative and 0 is definitely not cooperative. Error bars show the standard errors of the mean.

In the case where blue waits, he is seen as equally cooperative as yellow and the uncertainty is resolved. In the case where blue goes straight for the goal, the inference that blue is no cooperative is enforced. Finally, when they collide in the middle, the inference that both are competitive is made. However all it takes is for blue to wait in order for this inference to be reversed. Furthermore, the inference that yellow was not cooperative carried over into the future actions yellow made. These inferences were all captured by the model inferences.

In Figure 4-5 shows the social dilemma grids which also explicitly require coordination. This situation is modeled after a four-way stop where to successfully cooperate one agent must let the other agent pass. The model captures the variation in participants judgments

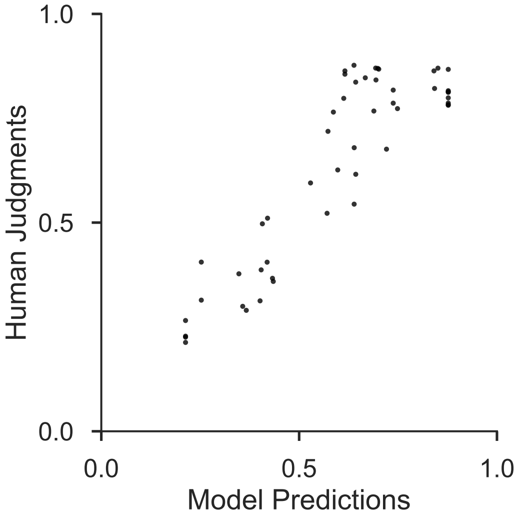


Figure 4-6: Model performance in the attribution experiment (R=0.9)

quite well but the overestimates cooperation when the agents collide in the middle. This could have been driven by carrying too much of the prior over from the situation where they both wait but also could have been due to using a higher order model such as asking whether or not the two agents are acting as a coordinated group. Across these two types of stimuli the model fit quite well which is quantified in Figure 4-6.

4.4 Discussion

In this work we developed a hierarchical model of social planning to understand how humans coordinate their low-level action plans to realize high-level strategic goals such as cooperation and competition. We formalize cooperation and competition as abstract planning procedures over low-level actions. Both model-based and model-free learning can create social norms which facilitate robust and stable coordination. One of our main contributions is formalizing how cooperative norms can make cooperation more robust across environments, a key step for long-lasting collaborative endeavors. While we only had space to show a subset of our full results, we are currently looking at how agents use these planning programs and the norms that they learn to generalize cooperation to completely new environments with the same partner. We will also use these models to study how observers

	Cooperate	Compete
Cooperate	$V^{\mathtt{Dec}}_{\mathtt{Dec}}(s), V^{\mathtt{Dec}}_{\mathtt{Dec}}(s)$	$V^{\mathtt{Dec}}_{\mathtt{BR}}(s), V^{\mathtt{BR}}_{\mathtt{Dec}}(s)$
Compete	$V^{\mathtt{BR}}_{\mathtt{Dec}}(s), V^{\mathtt{Dec}}_{\mathtt{BR}}(s)$	$V^{\mathtt{BR}}_{\mathtt{BR}}(s), V^{\mathtt{BR}}_{\mathtt{BR}}(s)$

Figure 4-7: Abstracted game using social planning. $V^{\mathtt{Dec}}_{\mathtt{Dec}}(s)$ is the value function when both players are using the decentralized cooperative plan, $V^{\mathtt{BR}}_{\mathtt{Dec}}(s)$ is the value function when the first player best responds to other playing the cooperative plan, and $V^{\mathtt{BR}}_{\mathtt{BR}}(s)$ is the value function when both players best respond against each other.

attribute cooperative and competitive intentions to other agents.

These hierarchical planning programs allow agents to reduce the infinite sized matrix show in Figure 4-2 down to the more compact representation shown in Figure 4-7. Furthermore, these functions allow us to categorize games according to different abstract features. For instance, to describe the coordination challenge inherent in a game we develop a metric called the "cost of autonomy:" $V^{\mathtt{Central}}(s) - V^{\mathtt{Dec}}_{\mathtt{Dec}}(s)$ which is the difference between what an agent could receive in expectation if they were actually controlled by a "we-agent" and what they will get in expectation running a decentralized plan. Another interesting metric is the "temptation:" $V^{\mathtt{BR}}_{\mathtt{Dec}}(s) - V^{\mathtt{Dec}}_{\mathtt{Dec}}(s)$ which is the difference between what an agent will get in expectation when executing the decentralized cooperative plan and what they will get from best responding to that plan. This is a measurement of the degree to which using the competitive planner is favored in the short term. In future work we will use these metrics to characterize different the different spatial grids.

While cooperation and coordination are often studied as separate phenomenon cooperation often requires complex coordination. Consider two different teams of scientists. In one pair, one designs the experiment and the other collects the data, in the other pair, the pair sit in a room brainstorming new ideas together. Or consider the difference between two chefs where one cooks a main course and one cooks a dessert and two chefs who work together making just a single dish. The difference between the former over the later is the complex "meshing" of subplans. In the former cases, the the initial coordination is sufficient for a cooperative outcome, as long as they both don't choose to carry out the same task (two desserts or two designed experiments), the cooperation is likely to be successful. In the later cases, cooperation requires continuous coordination, mutual responsiveness and intention inference, much like a jazz duet. Consider the following definitions which describe the criteria for a joint intention or we-intention:

Definition 1. We intend to J if and only if (M. E. Bratman, 1993, 2014):

1. (a) I intend that we J and (b) you intend that we J.

2. I intend that we J in accordance with and because of 1a, 1b and meshing subplans of 1a and 1b; you intend that we J in accordance with and because of 1a, 1b, and meshing subplans of 1a and 1b.

3. 1 and 2 are common knowledge between us.

In future work, we hope to describe how our computational model described here can capture some aspects of the philosophical notion of we-agency.

One interesting feature of the model is how an asymmetric w in the cooperative planner can break symmetries making successful coordination more likely. In future work we'd like to explore how priors on this parameter in social hierarchies might enable more effective teamwork e.g., boss-employee relations (Galinsky & Schweitzer, 2015). Finally, in our current paradigm, the desires of all agents are common knowledge. Investigating environments that require jointly inferring the goals of others and the plan needed to help realize a cooperative outcome will be examined in future work.

In future work we will extend these models to even more realistic and complex domains. One promising direction is to study how the structure of these algorithms might allow for cooperation in real-time games with complex objects and physics. In these complex multi-agent interactions, people often cooperate `to` compete, coordinating cooperation with some agents in order to better compete against others. Finally, we can use these games to study how primitive but distinctively human forms of communication such as gesture and pointing can help initiate, sustain, and structure complex collaborations. By grounding strategic social reasoning in a theory of planning we can begin to investigate the mechanisms of joint intentionality and how these joint intentions enable the scale and scope of human cooperative behavior (Tomasello, 2014).

Chapter 5

Non-parametric Bayesian Inference of Strategies

5.1 Introduction

In strategic settings, predicting the actions of other players is essential for both cooperative and competitive intelligent behavior. Models of how people reason about and infer the strategies of others can give insights into the cognitive systems used by humans in interactive strategics contexts. Repeated games are a key example: players must infer the strategies of others based on their previous interactions in order to achieve their cooperative or competitive goals.

Repeated games offer significantly more strategic possibilities than one-shot games. In this work we will use the repeated two-player prisoner's dilemma as an example to demonstrate our approach, although our model is general in the underlying stage game and in the number of players. We first briefly describe the one-shot prisoners dilemma: two players simultaneously choose to either 'cooperate' (C) or 'defect' (D). Based on their joint selection of actions, they obtain utility according to the payoff matrix in Figure 5-1. When the game is played only once, there is only a single Nash equilibrium: both players defect. Thus the prisoner's dilemma represents a simplified social dilemma, both players would prefer to receive the pareto-optimal outcome (C, C) but only (D, D) is the equilibrium outcome.

		Player 2	
		C	D
Player 1	C	a,a	12,50
	D	50,12	25,25

Figure 5-1: Payoff matrix for a two-player prisoners dilemma. The value of a sets the payoff of joint cooperation and must be $25 < a < 50$.

When the prisoner's dilemma is repeated an indefinite number of times between the same two players, the cooperative outcome can be rationally sustained as dictated by the folk theorem (Fudenberg & Maskin, 1986). Unlike one-shot games where the space of strategies is just a measure over the action space, rational players in repeated games condition their actions on previous outcomes. This results in an exponential growth in the number of strategies as the repeated interaction continues. As a result of this exponential growth, learning these strategies from data seems like an intractable task. Each additional round requires conditioning on greater and greater amounts of data. This complexity is sometimes called the *curse of history* (Pineau, Gordon, Thrun, et al., 2003).

To succinctly represent strategies with behavioral significance, theorists have turned to a model of bounded computation – the finite state transducer (FST), a type of automaton which compactly represents strategies using only limited memory Carmel & Markovitch, 1996, and Rubinstein, 1986. In the prisoners dilemma, a conditional cooperation strategy called Tit-for-Tat (TFT), which starts off playing (C) and then copies the previous move of the other player, can be compactly represented using FSTs. These simple strategies have significance for the evolution of cooperation, understanding human behavior, and designing self-regulating cooperative systems (Axelrod & Hamilton, 1981; Kleiman-Weiner et al., 2016; Littman & Stone, 2005). Below we show an example interaction in the two-player repeated prisoners dilemma where player 1 (P1) is using the TFT strategy:

P1: $CCCDCCCDDDC\ldots$

P2: $CCDCCCDDDCC\ldots$

Simple strategies like TFT are often enriched by considering forgiving variants which return to cooperation after a string of mutual defections or a vindictive TFT which defects multiple times after a defection regardless of whether or not the other player returns to cooperation (M. A. Nowak & Sigmund, 1992; Zagorsky, Reiter, Chatterjee, & Nowak, 2013). Developing strategies for repeated games in terms of FST enables theorists to capture and study their intuitions about behavior in a formal model. Using FSTs to represent strategies, one maps the *curse of history* of strategy inference (where the number of strategies grows exponentially in previous interactions) to a search over the space of possible FSTs. However, there are still *a priori* an infinite number of possible automata one must consider.

So how might people infer strategies from the behavior of other players? How do theorists generate new candidate FSTs for study from this infinite space? In this work we develop a Bayesian model for strategy inference. The problem of strategy inference can be posed probabilistically as finding $P(\text{strategy}|\text{data})$ i.e., given data from an interaction between players, finding the probability of each strategy (as represented by an FST). Using Bayes rule we can write the posterior distribution in terms of the data likelihood and a strategy prior:

$$P(\text{strategy}|\text{data}) \propto P(\text{data}|\text{strategy})P(\text{strategy})$$

The core of this work is to formalize the pieces of this relationship and to propose an algorithm for inference. $P(\text{data}|\text{strategy})$ is the probability that an FST could have generated a specific sequence of behavioral data. For deterministic FST, this probability distribution is a delta function. However, in reality, behavior is likely to be ‘noisy’ i.e., selected play may not coincide exactly with the action prescribed or intended by the strategy. If we assume probabilistic errors, any FST can generate a sequence of data but with varying probability. The real challenge of inference comes from specifying $P(\text{strategy})$, the prior distribution over possible strategies. Since strategies are represented as FSTs and we want to consider strategies of arbitrary complexity, this is equivalent to specifying a distribution over all possible FSTs.

Solving this inference problem has key implications for learning equilibrium strategies. Under some general assumptions, rational learning leads to Nash Equilibria in infinitely repeated games (Kalai & Lehrer, 1993). If each player assigns positive probability to all remaining possible opponent strategies that can occur within future play given past observations, then Bayesian updating will lead in the long run to accurate predictions about future play of the game. Bayesian updating is essentially a process of eliminating 'impossible' strategies, and selecting the most probable strategies from the remaining possible choices. The prior plays a key role, in order for guarantees on rational learning to hold, a learner must correctly assign positive probability over all possible remaining strategies.

One solution common in experimental game theory is to put a uniform distribution over a hand-selected subset of strategies ((Bó, 2005; Bó & Fréchette, 2011; Blonski, Ockenfels, & Spagnolo, 2011). However this approach only allows for one to estimate the relative likelihood of strategies that are specifically hypothesized *a priori*, preventing the discovery of novel strategies for commonly studied games. Furthermore, this method is not robust for games that have not been analytically analyzed. For instance, Bó & Fréchette, 2011 conducted experiments and found that Tit-for-Tat and Always-Defect account for more than 80% of played strategies in Repeated Prisoner's Dilemma games, but later work pointed out that a lesser known strategy of equal complexity called Semi-Grim can better account for their data (Breitmoser, 2015). Since Semi-Grim was not in the authors' original prior hypothesis space, it could not be inferred. Likewise, inexperienced players faced with a strategic situation need a robust way of inferring the strategies of other players such that given sufficient evidence, the correct strategy will be inferred.

Besides using a uniform prior over a finite hypothesis space, another approach for predicting behavior is based on learning methods such as fictitious play (Fudenberg & Levine, 1998). Under this framework, each player best responds to the other player based on the empirical frequency of the other player's actions. While these methods can be powerful at predicting behavior, they do not infer a model of the other players' strategies. Thus, while reinforcement learning methods can learn the statistical likelihood of certain actions, they do not learn a *causal* model (like a FST) of other players. These methods are less likely to generalize across games or predict the behavior of others in rare situations.

Here, we present a novel non-parametric Bayesian model for strategy inference in repeated games. We develop a new prior over strategies based on the Hierarchical Dirichlet

Process (HDP) (Teh, Jordan, Beal, & Blei, 2006). The model is non-parametric in the number of states in an FST and implicitly represents the infinite space of possible FST. We derive a Gibbs sampler for efficient inference in this model which successfully infers the actual strategy in simulated interactions. Our model predicts the correct FST even under noisy conditions and can be used to investigate human strategies from behavioral data. Our main contribution is to bring powerful tools from statistical machine learning to the study of strategic behavior. To our knowledge this is the first application of the HDP in game theory and the analysis of human strategic behavior in games. This model is a step towards developing computational agents with social intelligence that can predict the behavior of others in strategic settings.

5.2 Model

We first describe the finite state transducer (FST) formally and describe its relation to a hidden Markov model (HMM). This relation allows us to leverage a suite of tools from probabilistic graphical models for inferring FSTs. We review the hierarchical Dirichlet process (HDP) and the HDP-HMM and extend these models to represent strategies. We call this new model the HDP-FST. The HDP-FST is a generalization of the HDP-HMM and can be used to represent and infer strategies in repeated games.

5.2.1 FST and HMM

An FST is a bounded model of computation which is capable of representing strategies in infinitely repeated games (Rubinstein, 1986). Formally, for player i an FST is a tuple $\langle S_i, \boldsymbol{O}, \boldsymbol{y}, \boldsymbol{\pi}, \phi, \mathrm{F} \rangle$ namely, a finite set of states $S_i = \{s_1, \ldots, s_n\}$, a finite set of input symbols $\boldsymbol{O} = \{o_1, \ldots, o_n\}$, a finite set of emission symbols $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, and a transition relation $\boldsymbol{\pi}$ where $\pi_{ij} = Pr(s_{t+1} = j | s_t = i, o_t \in \boldsymbol{O})$, where o_t is the input observed at time t. F characterizes the distribution for emissions at each state $s_i \in S_i$, where ϕ_{s_i} parameterizes the emission y_i such that $y_i | s_i \sim \mathrm{F}(\phi_{s_i})$.

FSTs represent strategies as "if-then" computations. Given two players i and j: *if* player j previously played action o_j, *then* player i transitions to a particular state s and performs action $y_i|s$ at the next iteration of the game. Thus FSTs take in a set of inputs and for each input, update their internal state and produce an emission. In the context of strategies, the emissions of an FST correspond to actions.

For illustration, two FSTs (TFT and Semi-Grim) are reproduced in Figure 5-2, with $S_i = \{s_C, s_D\}$, $\boldsymbol{O} = \{C, D\}$, $\mathbf{y} = \{C, D\}$, $f_{\phi_{s_C}}(C) = 1$ and $f_{\phi_{s_D}}(D) = 1$, where $f_{\phi_{s_C}}$, $f_{\phi_{s_D}}$ are the pdf of $F(\phi_{s_C})$ and $F(\phi_{s_D})$ respectively. Since the input symbols in games come from other strategic players we will use $\mathbf{y}_{-i} = \boldsymbol{O}$ where $-i$ are all the players except i. The starting state of an FST is s_0.

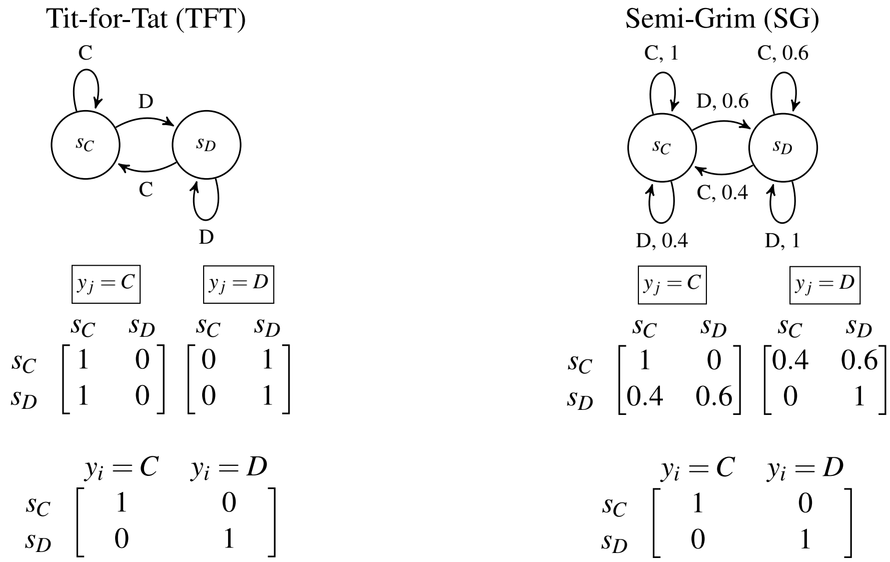

Tit-for-Tat (TFT)
C
D
C
D
s_C
s_D
$y_j = C$
$y_j = D$
$\begin{array}{c|cc} & s_C & s_D \\ s_C & 1 & 0 \\ s_D & 1 & 0 \end{array}$
$\begin{array}{c|cc} & s_C & s_D \\ s_C & 0 & 1 \\ s_D & 0 & 1 \end{array}$
$\begin{array}{c|cc} & y_i = C & y_i = D \\ s_C & 1 & 0 \\ s_D & 0 & 1 \end{array}$
Semi-Grim (SG)
C, 1
C, 0.6
D, 0.6
C, 0.4
D, 0.4
D, 1
s_C
s_D
$y_j = C$
$y_j = D$
$\begin{array}{c|cc} & s_C & s_D \\ s_C & 1 & 0 \\ s_D & 0.4 & 0.6 \end{array}$
$\begin{array}{c|cc} & s_C & s_D \\ s_C & 0.4 & 0.6 \\ s_D & 0 & 1 \end{array}$
$\begin{array}{c|cc} & y_i = C & y_i = D \\ s_C & 1 & 0 \\ s_D & 0 & 1 \end{array}$

Figure 5-2: Representations of strategies for the two-player repeated prisoners dilemma. (top) FST representation. Arrows show transitions between states given the actions of the other players. (middle) Transition matrices between the state at time t (rows) and $t+1$ (columns), one for each action available to the other player. Each entry gives the probability of transitioning between states. The box above each matrix specifies the other player's action at time t. (bottom) Emission matrix which probabilistically maps from a state (rows) to action (columns).

Let $s_{i,t} \in S_i$ be the state of player i and $y_{i,t} \in y_i$ be the action taken by player i at time t. Using TFT as an example, $s_{i,t} = s_C$ and $y_{i,t} = C$. If the action by player j at t, $y_{j,t} = C$, then $s_{i,t+1} = s_C$ and $y_{i,t+1} = C$; otherwise if $y_{j,t} = D$, then $s_{i,t+1} = s_D$ and $y_{i,t+1} = D$ from $\boldsymbol{\pi}$. Thus i plays in round $t+1$ the action that the j played in round t.

The strategy Grim-Trigger plays cooperate until a defection is played and then defects forever. Semi-Grim is a more forgiving version of the Grim-Trigger that "forgives" defection and tries to resume cooperation i.e., even if $\exists k \in \{1, \ldots, T\}$ s.t. $y_{j,k} = D$, there is a small probability for s_i to return to s_C if $y_{j,t} = C$. Following the example in Figure 5-2, if $y_{j,t} = D$ and $s_{i,t} = s_C$, then $Pr(s_{i,t+1} = s_C) = 0.4$. Thus there is some probability that i remains in a cooperative state even if j defects. Similarly, if $s_{i,t} = s_D$ and $y_{j,t} = C$, then $Pr(s_{i,t+1} = s_C) = 0.4$ i.e., there is some probability of transitioning back to the cooperating state if cooperate was played by j.

In order to infer FST from data, we first describe the relationship between an FST and the HMM, a common probabilistic model for analysis of sequential data such as language or DNA. An HMM is a doubly-stochastic Markov chain defined as a tuple $\langle \boldsymbol{s}, \boldsymbol{\pi}, \boldsymbol{y}, \phi, \mathrm{F} \rangle$; where $\boldsymbol{s} = (s_1, s_2, \ldots, s_T)$ represents a sequence of states linked by a transition matrix $\boldsymbol{\pi}$, where $\pi_{ij} = p(s_{t+1} = j | s_t = i)$, with $\pi_{0i} = p(s_1 = i)$. Corresponding to each state in the model is a parallel sequence of observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ with y_t drawn conditionally dependent only on s_t. For each state $s_t \in \{1, \ldots, K\}$ there is a parameter ϕ_{s_t} that reflects the likelihood of the observation at that state: $y_t | s_t \sim F(\phi_{s_t})$.

The difference between an HMM and an FST is that an HMM does not condition on the observed actions made by player j. Conditional on an opponent's action at $t-1$, both models are representationally identical – a set of transition matrices in the FST becomes a single transition matrix like the HMM. This implies that the HMM can be written as a limiting case of the FST. By assuming that player j's action fully specifies the transition probability between s_{t-1} and s_t i.e., each row of the transition matrix is conditionally independent, then the HMM can be augmented into an FST by making the states of the HMM dependent on the other players' actions. The equivalence between these augmented HMMs and FSTs has been formally proven and has been applied for use in speech recognition (Kempe, 1997; Mohri, Pereira, & Riley, 2002).

While the FST formalism can represent specific strategies, it doesn't provide an algorithm or mechanism for enumerating or representing a hypothesis space of strategies. Consider an example sequence of plays in the infinitely repeated prisoner's dilemma where the observed actions are [(D,D), (D,C), ...]. Player 1's strategy could be Always-Defect, which is a one-state FST or could be TFT, a two-state FST, or even a three-state FST where player 1 begins with D, and defects until two iterations of C is observed from player 2, then plays C. This kind of reasoning can generate an infinite space of strategies if the number of states in the FST are not restricted. How do we represent this space formally and apply it tractably for inference?

5.2.2 HDP-HMM

We now develop a new model for strategy inference based on the correspondence between the HMM and FSTs. This model is based on the HDP-HMM, a non-parametric extension of the HMMs (Teh et al., 2006). This is the first time to our knowledge of Bayesian non-parametric models applied to a game theoretic contexts. We first review Bayesian non-parametric models in general, focusing on Dirichlet Processes. Then we describe how Dirichlet Processes can be generalized with the Hierarchical Dirichlet Process (HDP) and how these tools are used for sequence modeling with the HDP-HMM.

We first introduce the Dirichlet Process (DP). A DP is a generalized Dirichlet distribution (the conjugate prior of a multinomial distribution), but may contain an infinite number of elements. The DP is commonly used to describe a prior over the distribution of random variables and is parameterized by a base distribution H, and a concentration parameter α, where $\alpha > 0$. For instance, consider the following DP

$$\begin{aligned} G &\sim DP(\alpha, H) \\ H &\sim N(\mu, \sigma^2) \end{aligned} \tag{5.1}$$

where the base distribution (H) is a normal distribution with mean μ and variance σ^2. Draws from the DP, G, would have the same support as H, with one important difference - all draws from the DP are discrete. While H is continuous, implying that the probability that any two samples are equal is 0, this is not the case for G. See Figure 5-3 below for a graphical illustration.

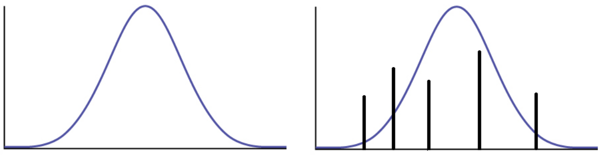


Figure 5-3: (left) A normal distribution $H \sim N(\mu, \sigma^2)$ and (right) $G \sim DP(H, \alpha)$, with H interposed for reference. G is a probability distribution that "looks like" H, but whose distribution is discrete.

Stick-Breaking Process

The DP can also be represented by a stick-breaking process. This formalism directly reveals its discrete nature (Sethuraman, 1994). For $k = 1, 2, \ldots$, let:

$$\phi_k \sim H \qquad \beta'_k \sim Beta(1, \alpha) \qquad \beta_k = \beta'_k \Pi_{l=1}^{k-1}(1 - \beta'_k) \tag{5.2}$$

Then the random measure defined by $G = \sum_{k=1}^{K} \beta_k \delta_{\phi_k}$ is with probability one equal to a sample from $DP(\alpha, H)$, where δ_ϕ is a probability measure concentrated at ϕ. The construction of β_1, β_2 can be also thought of starting off with a stick of length 1, and we break it off at $\beta_1 \sim Beta(1, \alpha)$, and recursively break the remaining portion at β_2, β_3 and so on. This process is also called GEM(α) after Griffiths, Engen and McCloskey, where α refers to the same concentration parameter as in the DP, and $\alpha > 0$.

Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP) is a set of DPs that are coupled together with a random base measure that is itself a DP. The HDP was developed to apply non-parametric methods to the problem of clustering grouped data. Data can be subdivided into different groups, and within each group there can be clusters that capture latent structure within that group. These models have been used in machine learning to cluster and classify data such as documents and genetic data (Beal, Ghahramani, & Rasmussen, 2002; Blei, Ng, & Jordan, 2003; Gabriel et al., 2002; Wood, Gasthaus, Archambeau, James, & Teh, 2011).

There are many similarities between these problems and the challenge of inferring strategies. As we have shown, strategies can be represented in terms of states, transition probabilities, and emission matrices. The clusters in strategy inference refer to the distinct states in the FST. Given a sequence of observed actions by player i (e.g., ‘C’ or ‘D’), and the corresponding history of actions by player j, we want to identify the underlying and distinct states in i that produced these actions and the transitions between these states. However, in order to consider all possible strategies, we would have to consider and conduct inference for an infinite number of transition parameters and states, which is not a computationally feasible process.

Next, we need to integrate out the infinite number of transition parameters and represent the process with a finite number of indicator variables. In the HDP there is a natural bias towards using already existing transitions proportional to their prior usage (“rich get richer”). This implies that the latent state sequence ($\boldsymbol{s_i}$) produced by the FST that we observe are *typical trajectories* (Beal et al., 2002) – which results in a set of strategies biased towards those of lower complexity (as measured by the number of states in the FST) that most resemble the actual strategy.

We now formally describe the HDP. First, we define a global vector β and local vectors π_k in the method shown below.

$$\begin{gathered}\beta \sim \text{Dirichlet}(\gamma/K, \ldots, \gamma/K) \\ \pi_k | \beta \sim \text{DP}(\alpha, \beta) \qquad \phi_k \sim H\end{gathered} \tag{5.3}$$

where π_k represents the transition probabilities out of state k, and ϕ_k parameterizes the distribution of emissions at each state k, drawn from a base distribution H. Since each $\pi_k \sim DP(\alpha, \beta)$, the states (within each FST) that the transition matrices refer to are *shared* as each DP is drawn from the same β, itself a *discrete* distribution obtained from the global draw parameterized with γ and K. Each atom in β hence represents the prior mean for transition probabilities leading into state k.

The ratio γ/K determines the sparsity of β. For instance, if $\gamma/K \ll 1$, the mass of β will be highly concentrated in just a few components. When $\gamma/K \to \infty$, the mass will gradually be equally dispersed across all the components. As $K \to \infty$, the prior in equation (5.3) becomes an HDP. Each π_k has concentration parameter α that determines deviation from the mean. This sharing is shown graphically in Figure 5-4. Since the DP draw of the base measure is necessarily discrete, subsequent draws will be drawn from the same discrete distribution.

Similar to the DP, the HDP can also be constructed via a stick-breaking process (Teh et al., 2006; Van Gael, Saatci, Teh, & Ghahramani, 2008):

$$\beta \sim \mathrm{GEM}(\gamma), \qquad \pi_k|\beta \sim \mathrm{DP}(\alpha,\beta), \qquad \phi_k \sim H$$
$$s_{t+1}|s_t \sim \mathrm{Multinomial}(\pi_{s_t}), \qquad y_t|s_t \sim \mathrm{F}(\phi_{s_t}) \tag{5.4}$$

The graphical model for the HDP-HMM is shown in Figure 5-5.

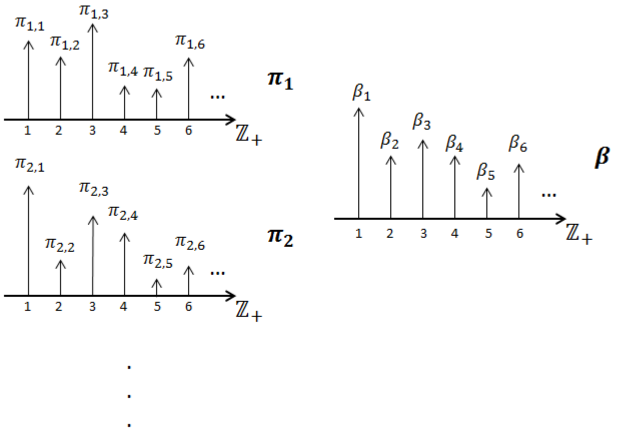


Figure 5-4: The sparsity of β is shared. (right) An example β. (left) Each π_i shares the same atoms as β and $\pi_{i,k}$ has β_k as its expected value.

5.2.3 HDP-FST

Just as we showed the relation between the FST and the HMM, we will now show that the HDP-HMM is a limiting case of a more general class of models which we call the HDP-FST. Let $\boldsymbol{y_i} = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})$ be the history of actions for player i where each $y_{i,t}$ is the action player i took at time t. $\boldsymbol{y_j}$ are the history of actions taken by player j which are observed by player i. For now, we restrict analysis to one additional player, but the model is general any finite number of players.

In an HMM, the probability distribution of each subsequent state is dependent only on

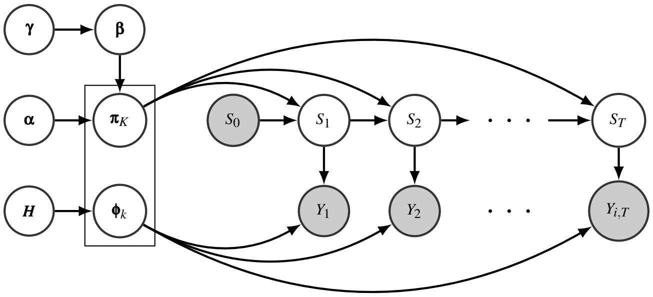


Figure 5-5: Graphical model for the HDP-HMM. The four 'global' parameters $(\gamma, \alpha, H, \beta)$ generate π_K and ϕ_k, the two parameters that define a FST. The remainder of the figure follow a typical HMM formalism, with the emission $Y_.$ being a function of the state $s_.$ and ϕ_k. Each state follows the Markov property, i.e. the probability distribution of the current state depends only on the previous state.

the previous state. However, in a FST each state is dependent on both the current state and the observed actions of the other player. Adding this additional dependency to the HDP-HMM turns it into the HDP-FST. Let the state sequence of player i be $\boldsymbol{s_i} = (s_{i,1}, \ldots, s_{i,T})$. Player i observes player j's action $y_{j,t}$ at time t and conditions $s_{i,t+1}$ on both the previous state $s_{i,t}$ and the previous action played by player j, $y_{j,t}$. Thus like an FST, conditional on s_t, the HDP-FST only needs to know the other player's previous action $y_{j,t}$ and not their full sequence of actions $\boldsymbol{y_j}$.

Heterogeneity between different people could be captured through the hyperparameters, β, ϕ and H which are now subscripted by i, but retain their interpretation from equation (5.4). The HDP-FST is formally:

$$\begin{aligned} \beta_i \sim \text{GEM}(\gamma_i), \quad \pi_{i,k,y_j}|\beta_i \sim \text{DP}(\alpha_i, \beta_i) \\ \phi_{i,k} \sim H_i \quad s_{i,t+1}|s_{i,t}, y_{j,t} \sim \text{Multinomial}(\pi_{i,s_{i,t},y_{j,t}}) \\ y_{i,t}|s_{i,t} \sim F(\phi_{i,s_{i,t}}) \end{aligned} \tag{5.5}$$

The key difference between equation (5.5) and equation (5.4) is that π_i additionally depends on y_j. This is comparable to how an HMM can be augmented into an FST. Intuitively, consider a partition of π_i into $|y_j|$ different $k \times k$ matrices, where $|y_j|$ is the number of unique actions available to player j, and when $|y_j| = 1$, then the HDP-FST is equivalent to the HDP-HMM. The interpretation of the hyperparameters $\beta_i, \alpha_i, \gamma_i$ is unchanged.

Figure 5-6 shows the graphical model for the HDP-FST in the case of two players.

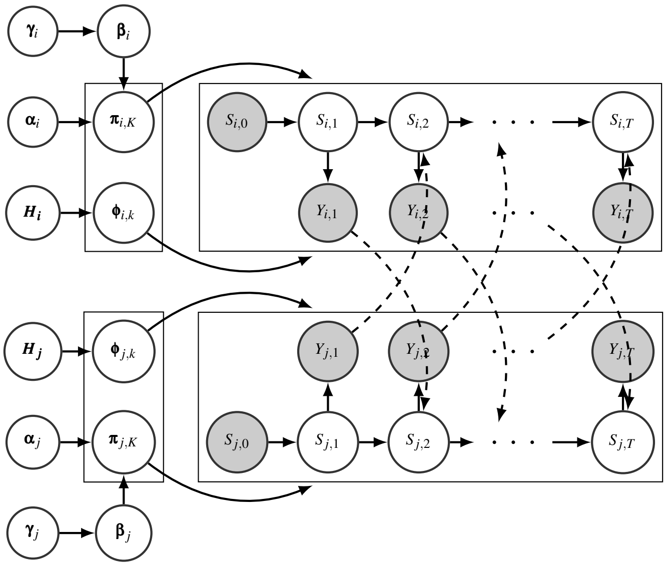


Figure 5-6: HDP-FST graphical model. Some arrows from the hyperparameters to the states $s_{i,t}$ and actions are omitted for clarity. The dotted arrows represent the additional conditional dependency added from the HDP-HMM model, where each $s_{i,t}$ is conditioned on $y_{j \neq i,t-1}$, $s_{i,t-1}$ and π_i.

Note that each player is shown with a different set of hyperparameters. These hyperparameters could be different for different players if we believed that the players differed in some capacity. A larger γ for a particular player could correspond to believing that a player is a priori more likely to play a smaller FST. Similarly, choosing a smaller γ corresponds to a higher prior probability on strategies with a large number of states allowing for larger FSTs. While the model allows for this variation, in this work we use the same parameters and base distribution for all analyses. Hence for the remainder of the paper:

$$\gamma_i = \gamma_j = \gamma \qquad \alpha_i = \alpha_j = \alpha \qquad H_i = H_j = H$$

5.2.4 Inference

Having described a prior over strategies using HDP-FST, we now describe how to make inference over this hypothesis space tractable using a Gibbs sampler. The Gibbs sampler is a commonly used method for drawing samples from a distribution that cannot be calculated analytically. This method is relevant because the exact Bayesian inference for the model is intractable as the number of states K is infinite, which means we cannot apply a generic forward-backward algorithm as one would commonly do if the K was known in advance. Since the distribution is over strategies, each sample is a specific FST and by running the Gibbs sampler for many iterations we can draw enough samples to approximate the posterior distribution.

We now describe the Gibbs sampler for inference in the HDP-FST. Gibbs sampling works by sampling each variable while conditioning on the values of all the other variables in the distribution (Murphy, 2012). Our Gibbs sampler builds on the direct sampling mechanism presented in Teh et al., 2006, and we reference Van Gael et al., 2008's description of sampling for the HDP-HMM. Given that the states are exchangeable, we can analytically marginalize out the latent variables π_i, ϕ_i from equation (5.5). Thus sampling will only involve the latent state sequence $\boldsymbol{s_i}$ and the DP parameter β_i. In order to resample $s_{i,t}$, we need to calculate the following conditional probabilities:

$$\begin{aligned}&p(s_{i,t}, s_{j,t} | s_{i,-t}, s_{j,-t}, y_{i,t}, y_{j,t}, y_{i,t-1}, y_{j,t-1}, \beta_i, \beta_j, \alpha, H)\\ &\propto p(y_{i,t}|s_{i,t}, H) \cdot p(y_{j,t}|s_{j,t}, H) \cdot p(s_{i,t}, s_{j,t}|s_{i,-t}, s_{j,-t}, y_{i,t-1}, y_{j,t-1}, \beta_i, \beta_j, \alpha) \end{aligned} \tag{5.6}$$

where $s_{i,-t}$ refers to the sequence of states s_i excluding $s_{i,t}$.

However, the structure of conditional independence in the HDP-FST can simplify this equation and allow for more efficient sampling. The conditional likelihood of $(y_{i,t}, y_{j,t})$ given the states $s_{i,t}$ and $s_{j,t}$, actions $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$, and base distribution H is easy to compute as each $y_{i,t}$ and $y_{j,t}$ is only dependent on their respective states at time t. Further, if the base distribution H and the likelihood F from equation (5.5) are conjugate we can analytically update this portion of the likelihood. Given that the state space for these strategies is discrete, the conjugate multinomial-Dirichlet distribution is appropriate and greatly simplifies inference.

Furthermore, we can use the independence structure of the model to avoid having to sample from $\{\boldsymbol{s_i}, \boldsymbol{s_j}\}$ jointly. Because of the Markov property of the model, each $s_{i,t}$ is conditionally independent of all $\boldsymbol{s_j}$ given $s_{i,t-1}$, $s_{i,t+1}$, $y_{j,t-1}$ and $y_{j,t}$, where $y_{j,t-1}$ and $y_{j,t}$

are observed. Therefore each sequence of states $\boldsymbol{s_i}$ can be sampled independently of the hidden states of all other players. This factors the model into independent components (one for each player) which can be treated separately during inference. For this reason, we can sample each player independently when resampling the latent state sequences. We can also further reduce $s_{i,-t}$ to $\{s_{i,t-1}, s_{i,t+1}\}$ given that $s_{i,t}$ is only dependent on player i's state one time period prior and one after. Using this simplification we can rewrite equation (5.6) as:

$$\begin{aligned} & p(s_{i,t}|s_{i,t-1}, s_{i,t+1}, y_{i,t}, y_{j,t-1}, y_{j,t}, \beta_i, \alpha, H) \\ & \propto p(y_{i,t}|s_{i,t}, H) \cdot p(s_{i,t}|s_{i,t-1}, s_{i,t+1}, y_{j,t-1}, y_{j,t}, \beta_i, \alpha) \end{aligned} \tag{5.7}$$

Now we describe the sampling mechanism for $p(s_{i,t}|s_{i,t-1}, s_{i,t+1}, y_{j,t-1}, y_{j,t}, \beta_i, \alpha)$. First, we remove the subscripts for s_i, β_i and observed j's actions y_j as it is clear which player we are referring to for each variable. At any time t, let $n_{l,m}$ be the total number of transitions from sampled states l to m, excluding time steps t and $t-1$, and let $n_{\cdot,l}$, $n_{l,\cdot}$ be the total number of transitions into and out of state l respectively. Let K be the current number of distinct states in $s_1, s_2, \ldots, s_{t-1}$.[1]

Because there is a Dirichlet prior on β, we can define the distribution of s_t generated

[1] Since we ignore the ordering of states in β, the K distinct states are labeled $1, \ldots, K$, and $K+1$ refers to a new state.

from a single transition π_l[2] as:

$$\begin{aligned}
p(s_t|s_{t-1}=l,\beta,\alpha) &= \int_{\pi_l} p(\pi_l, s_t|s_{t-1}=l,\beta,\alpha)d\pi_l \\
&= \int_{\pi_l} p(\pi_l|\beta,\alpha)p(s_t|s_{t-1}=l,\pi_l)d\pi_l \\
&= \int_{\pi_l} \frac{\Gamma(\sum_k \alpha\beta_k)}{\prod_k \Gamma(\alpha\beta_k)} p(s_t|s_{t-1}=l,\pi_l)d\pi_l \qquad (5.8) \\
&= \int_{\pi_l} \frac{\Gamma(\sum_k \alpha\beta_k)}{\prod_k \Gamma(\alpha\beta_k)} \prod_{k=1}^{K+1} \pi_{l,k}^{\alpha\beta_k - 1} \prod_{k=1}^{K+1} \pi_{l,k}^{n_{l,k}} d\pi_l \qquad (5.9) \\
&= \frac{\Gamma(\sum_k \alpha\beta_k)}{\prod_k \Gamma(\alpha\beta_k)} \frac{\prod_k \Gamma(\alpha\beta_k + n_{l,k})}{\Gamma(\sum_k \alpha\beta_k + n_{l,k})} \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{l,\cdot})} \prod_k \frac{\Gamma(\alpha\beta_k + n_{l,k})}{\Gamma(\alpha\beta_k)} \qquad (5.10)
\end{aligned}$$

where we use the fact that β has a dirichlet prior in equation (5.8).

We augment (5.10) with the observed y_{t-1} to find, for a single state $s_t = k$:

$$\begin{aligned}
p(s_t = k|s_{t-1}=l, y_{t-1}, \alpha, \beta) &= \frac{\Gamma(\alpha + n_{l,\cdot|y_{t-1}})}{\Gamma(\alpha + n_{l,\cdot|y_{t-1}} + 1)} \frac{\Gamma(\alpha\beta_k + n_{l,k|y_{t-1}} + 1)}{\Gamma(\alpha\beta_k + n_{l,k|y_{t-1}})} \\
&= \frac{\alpha\beta_k + n_{l,k|y_{t-1}}}{\alpha + n_{l,\cdot|y_{t-1}}} \qquad (5.11)
\end{aligned}$$

which states that the probability of $s_t = k$ given $s_{t-1} = l$ is proportional to the relative frequency of previous transitions from state l to k ($\frac{n_{l,k}}{n_{l,\cdot}}$) and smoothed by $\alpha\beta_k$, the prior over k.

Note that we can sample backwards t from $t+1$ as well. Since we observe y_t (i.e. action of player j at time t), we have:

$$p(s_t = k|s_{t+1} = m, y_t, \alpha, \beta) = \frac{\alpha\beta_k + n_{k,m|y_t}}{\alpha + n_{k,\cdot|y_t}} \qquad (5.12)$$

[2] recall that π_l refers to the transition probabilities from state l to all other states, and $\pi_{l,k}$ refers to the probability of transitioning from state l to state k

From (5.11) and (5.12), we have: $p(s_t = k|s_{t-1}, s_{t+1}, \beta, \alpha, y_{t-1}, y_t) \propto$

$$\begin{cases} (\alpha\beta_k + n_{s_{t-1},k|y_{t-1}})\dfrac{\alpha\beta_{s_{t+1}} + n_{k,s_{t+1}|y_t}}{\alpha + n_{k,\cdot|y_t}} & \text{for } k \leq K, k \neq s_{t-1} \\ (\alpha\beta_k + n_{s_{t-1},k|y_{t-1}})\dfrac{\alpha\beta_{s_{t+1}} + n_{k,s_{t+1}|y_t} + 1}{\alpha + n_{k,\cdot|y_t} + 1} & \text{for } k = s_{t-1} = s_{t+1} \\ (\alpha\beta_k + n_{s_{t-1},k|y_{t-1}})\dfrac{\alpha\beta_{s_{t+1}} + n_{k,s_{t+1}|y_t}}{\alpha + n_{k,\cdot|y_t} + 1} & \text{for } k = s_{t-1} \neq s_{t+1} \\ \alpha\beta_k\beta_{s_{t+1}} & \text{for } k = K+1 \end{cases}$$

where y_{t-1} and y_t refers to the observed actions of player j at $t-1$ and t. Using these equations for the conditional updates we can realize tractable inference in the HDP-FST using a Gibbs sampler.

5.3 Results

We empirically investigate the effectiveness of this model to infer FSTs under sparse and noisy observations of behavior. We apply the model to previously published data set of human behavior in the infinite discounted prisoners dilemma (Bó & Fréchette, 2011). To more precisely evaluate the model, we also analyze performance on a simulated data set where the ground truth strategies are known.

5.3.1 Behavioral Results

We used our model to analyze data from a previous human behavioral experiment run by Bó & Fréchette, 2011. Subjects were matched into dyads and played a discounted version of the prisoners dilemma i.e., the repeated interaction ended with constant probability after each round. In different interactions the value of a in the payoff matrix varied between 32 and 48 (see payoff matrix in Figure 5-1). Prior work predicts that as the cooperative outcomes becomes less attractive relative to mutual defection, subjects will be more likely to defect. We analyzed all dyadic interactions that lasted at least 10 rounds.

In Figure 5-7 we show the averaged posterior distribution across the 41 dyads that

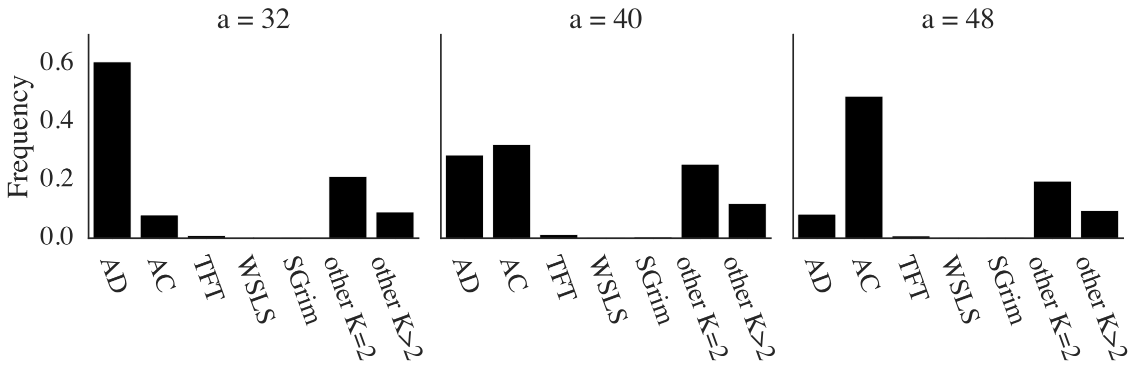
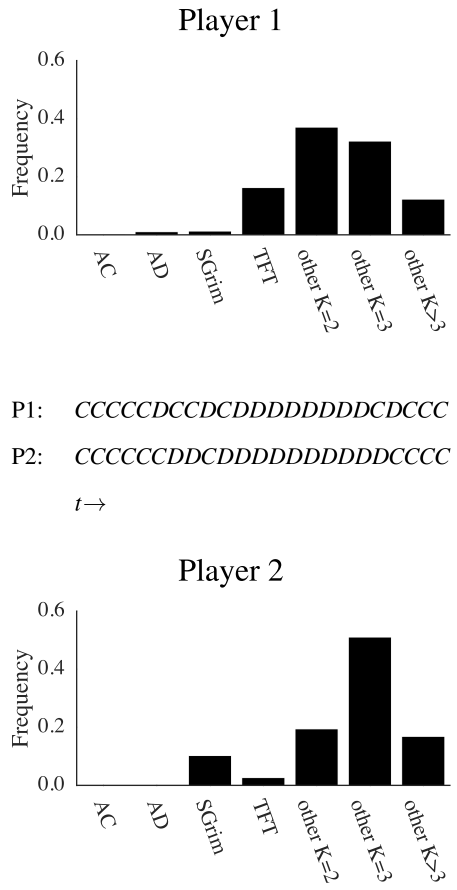


Figure 5-7: Posterior distribution over strategies averaged across all dyads that played at least 10 rounds of the prisoners dilemma. Strategies not commonly studied in the literature were clustered by their complexity where K is the number of states in the strategy. The value of a modulated the relative value of the cooperative outcome. Although most players are well described by simple one state strategies, there is a noticeable shift from most dyads playing always-defect (AD) when a was low (left) to most dyads playing always-cooperate (AC) when a is high (right).

lasted longer than 10 rounds for three values of a. When a took a low value the most common strategy inferred was Always-Defect (AD). As the value of a increased, the Always-Cooperate (AC) strategy was inferred with higher probability (with an intermediate balance of AC and AD when a took an intermediate value). We found relatively few instances of more complex strategies such as TFT and WSLS. This likely reflects a combination of the simplicity bias in the non-parametric prior since only a few of 41 dyads were longer than 20 rounds as well as the observation that many of the dyads played (C,C) or (D,D) for the entire interaction which while consistent with TFT and other more complex strategies does not provide for any additional predictive power over simpler strategies. Finally, we note that there was also a significant number of strategies the model discovered that have not been investigated in the literature, particularly those with $K > 2$ where K is the number of states.

We also analyzed the few long interactions in the data set that exceeded 20 rounds of repetition. We selected a dyad that seemed to have a fairly complex pattern of interaction and used the model to infer a distribution over strategies for just that dyad. Figure 5-8 shows the actions taken by the two players and the inferred distribution over strategies for each of the players. The model detected evidence of TFT for player 1 and Semi-Grim for



P1: *CCCCCDCCDCDDDDDDDDCDCCC*

P2: *CCCCCCDDCDDDDDDDDDDCCCC*

$t\rightarrow$

Figure 5-8: Posterior over strategies for two human players during a long interaction of 23 rounds. Of the strategies discussed in the literature, the model find some evidence for the play of Semi-Grim and TFT in both players. However, overall, both players' play are more consistent with larger and more complex FSTs.

player 2. As can be seen in the histograms both the two state FSTs were insufficient to capture the complexities of the interaction and most of the probability mass was on FSTs with three or more states. This suggests that people's actual strategies are more complex than the simple strategies of reciprocal cooperation (like TFT) predicted by current theory.

5.3.2 Simulated Results

Using the four FSTs listed in Figure 5-9: TFT, Win-Stay-Lose-Shift (WSLS), Tit-for-Two-Tats and semi-grim we generated 200 simulated dyadic interactions. Each of these simulations tests features of our model that have been challenging for previous approaches. We chose TFT and WSLS because they are of scientific significance (M. Nowak & Sigmund,

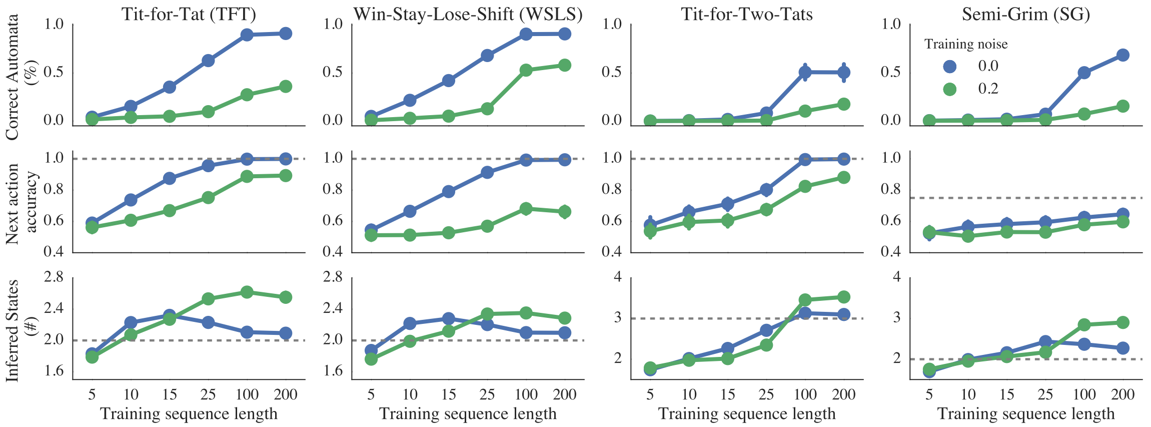


Figure 5-9: Simulated performance of inference with four different strategies (columns) that generated data while playing a random opponent. Each point is the average model performance of 200 runs. (Top) The posterior probability on the correct strategy. (Middle) The probability that the strategies contained in posterior distribution will correctly predict the next action of the true strategy. The dotted grey lines show the maximum possible performance. (Bottom) The average number of states in the posterior distribution of FSTs. The dotted grey lines show the actual number of states in the actual strategy.

1993), Tit-for-Two-Tats because it has more than two states and semi-grim because it has probabilistic transitions between states. Furthermore, we tested the algorithm when the observations were perfectly observed and also when 20% of the observations were corrupted by noise.

For all analyses presented here, we fixed $\alpha = \gamma = 0.5$ and used a symmetric prior on $H = [0.3, \ldots, 0.3]$. The first 200 samples of the Gibbs sampler were thrown away as burn in and the chain was thinned every 2 samples. We ran the sampler until we collected 500 posterior samples. While many of the FSTs described in the literature are those with deterministic transitions and emissions, the model is not restricted in this way and does not represent deterministic strategies any differently from probabilistic ones. Thus in order to compare the probabilistic output of strategies from the model, we round the transition and action matrices to their closest deterministic FST. Since two deterministic FSTs may be identical to each other by merely relabeling the states (which corresponds to permuting the transition and emission matrix), we clustered FSTs into functionally equivalent strategies by testing for isomorphisms in their graph. Using these two methods we were able to classify a sample of a probabilistic strategy from the model to a known strategy type (if one was known).

Figure 5-9 shows the results of this empirical analysis. We evaluated the inferences of the model on three metrics: its ability to infer the correct FST, whether or not it predicted the next action correctly, and the mean number of states in the distribution of sampled FSAs. Since the repeated prisoners dilemma has only two actions, chance guessing of the next action will result in 50% correct predictions. For semi-grim the best one can predict is only 75% due to stochasticity in the transitions. In contrast, the chance probability of predicting the correct FST is negligible. In no part of the model did we include any specification of the FSTs commonly studied in the literature but the model correctly infers them given only sparse and noisy observations.

The top row of Figure 5-9 shows the model's success in inferring the underlying strategy used to generate the interaction. Due to the simplicity bias inherited from the HDP prior, the simpler strategies (TFT and WSLS) are inferred correctly with less data. When trained on simulated behavior corrupted by noise (shown in green), predictive performance was impaired but still improved with more training data. Even when the model doesn't infer the correct FST with high accuracy it is still effective at predicting the next move. When the training sequence is short, there are many plausible FSTs that are consistent with the data.

Finally, we calculated the average number of states across the FSTs in the posterior sample as an approximate measure of the complexity of the inferred strategy. Given a very small amount of training data, the model mostly infers strategies of low complexity but as the amount of training data increases, the average number of states in the inferred sample grows. This feature, the ability to increase model complexity as the amount of data grows, comes from the non-parametric prior and balances against overfitting. With a medium amount of training data the number of FSTs considered grows considerably and even exceeds the actual number of states in the ground truth FST since there are many FSTs consistent with the data. However, as the training data increases, the model places most of its posterior mass on the correct strategy and the average complexity converges to the complexity of the true strategy. When the training signal is corrupted with noise, the complexity of the inferred FSTs exceeds that of the actual sequence since the model accounts for some of the stochasticity with extra model complexity to account for noisy actions.

With these simulated results we have shown the power of this model to infer the strategies used by players that play strategies described by a single FST out of an infinite space of possible strategies without ever enumerating that space. Our non-parametric model trades off model complexity with data fit and allows for the consideration of more complex models as the amount of data grows – in contrast to previous analyses which uses a finite hypothesis space.

5.4 Conclusions

In this work we developed the HDP-FST, a new non-parametric Bayesian model for the inference of strategies in repeated games. By extending the HDP, our model inherits many desirable properties of non-parametric models: (1) prior support over a hypothesis space that contains all possible FSTs without actually constructing this infinite space, (2) dynamically trades-off the complexity of the inferred model with model fit by biasing the posterior probabilities towards simpler strategies, and (3) allows for model complexity to grow with the data. We developed an efficient Gibbs sampler for conditional inference in this model. Using this inference scheme, we showed that from sparse and noisy observations of a dyadic interaction, it both infers the strategies and accurately predicts the expected next action. When applied to human data, the model inferred many strategies which have not been previously examined in the literature on repeated prisoners dilemma. While we focused on the infinitely repeated prisoners dilemma our model applies to any repeated game with a finite number of players and a finite action space.

In future work we would like to develop a beam sampler for these models which would allow for more efficient online inference (Van Gael et al., 2008). It may be possible to adapt these methods to account for nonstationarity in player's strategies by putting a lower weight on earlier actions. Since our approach to inference is probabilistic and causal it is possible to compose it with other probabilistic models allowing for richer multilevel analyses of human behavior that can explicitly model individual variation across subjects.

While our model of strategy inference considers all possible FSTs (with a bias towards simple strategies) it does not consider the strategic implications of the FSTs. Consider the following example:

P1: *CCCD*

P2: *CCCC*

where we want to infer the strategy for P2. While both AC and Tit-for-2-Tat are consistent with P2's play, we intuitively believe that Tit-For-2-Tat should be more likely, i.e., our intuitions about what strategies are most likely a priori may also take into account the strategic nature of those strategies, not just complexity. In future work we will investigate the way the prior over strategies itself might be modulated by payoffs:

$$P(\text{strategy}|\text{data},\text{payoffs}) \propto P(\text{data}|\text{strategy})P(\text{strategy}|\text{payoffs})$$

One possibility is that strategies which are not consistent with a best response could be assigned a lower or even zero probability in the prior. Another possibility is to weight a strategy's prior probability by an estimate of its expected payoff. The modulation of the prior by payoffs might itself be modulated by one's estimate of the strategic sophistication of one's opponent. For instance, if a player knew their opponent was very intelligent they might place a very low prior probability on that opponent using Always-Cooperate.

Understanding which strategies are "good" will likely require players that don't just infer strategies but also plan using them (Doshi-Velez, Wingate, Roy, & Tenenbaum, 2010; Kleiman-Weiner et al., 2016; Panella & Gmytrasiewicz, 2015). The combination of strategic inference with planning will be essential for developing intelligent agents that flexibly cooperate and compete.

Chapter 6

Fairness and the Inference of Reputation

6.1 Introduction

From the distribution of wealth across society to the distribution of dessert at the end of a dinner party, humans seem uniquely capable of enlarging the size of the pie and sharing it fairly (Tomasello, 2014). We make these decisions guided by normative principles such as efficiency, which says to maximize the total utility of the group and fairness, which says in part that distributions should be both equitable and impartial. We also use these principles intuitively when judging whether others' decisions are fair when considered from an impartial or objective perspective (Rawls, 1971; Nagel, 1986).

In the real world where resources aren't perfectly divisible, these principles can often come into conflict. It is well known that efficient allocations of resources are often inequitable and equitable allocations of resources are often inefficient – they leave some of the pie on the table. For example, if Alice has one apple and Bob has none and we take Alice's apple and throw it out, Alice and Bob are in a more equitable state but the total welfare (efficiency) is reduced. This is called inefficient equity. Even young children prefer inefficient equity: they prefer to destroy a resource rather than distribute it inequitably (P. R. Blake & McAuliffe, 2011; Shaw & Olson, 2012). Preferences for equity and efficiency are often captured quantitatively by directly deriving them from the outcomes. For instance, efficiency might correspond to the total or average outcome among a group of agents and inequity might correspond to the differences between the outcomes of different agents (Adams, 1965; Fehr & Schmidt, 1999).

While early work focused on whether a given outcome is perceived as fair (Adams, 1965; Fehr & Schmidt, 1999), there is now growing evidence that decision makers are sensitive to what their choice signals about themselves. Specifically, inequity created without showing partiality can be fair. If both Alice and Bob are equally deserving but there is only one apple, a decision maker might avoid giving it to either one in order to avoid an outcome that is neither equitable nor impartial. For instance, if the decision maker decided to give the apple to Alice an observer would infer that the decision maker is partial to Alice. However, if the decision maker can flip a coin or access another source of randomness and use the chance outcome to determine who should get the apple, the decision maker can create inequity but without worrying about others attributing partiality (Shaw & Olson, 2014; Choshen-Hillel, Shaw, & Caruso, 2015).

Both adults and children adjust their distributional preferences depending on whether they are the ones choosing or not. For instance, people are usually dissatisfied with receiving less than an equally worthy counterpart, but when they created the inequity themselves they were more likely to find this acceptable (Choshen-Hillel & Yaniv, 2011). Adults and children are willing to create inequity that disadvantages themselves but are less willing to create inequity that could be interpreted as favoritism or nepotistic preferences (Choshen-Hillel et al., 2015). These results are incompatible with explanations of social preferences that only consider an aversion to inequitable outcomes or other preferences that are directly derived from outcomes. Understanding how to combine these conflicting perspectives (efficiency vs. equity and equity vs. impartiality) is a challenge that we can address with computational modeling. Specifically, how might a flexible preference for these normative values be integrated together and flexibly applied?

Computationally, preferences like impartiality are significantly more sophisticated than just evaluating expected outcomes. We propose that an aversion to partiality is an aversion to having ones actions appear partial to others. Thus to evaluate whether an action will appear partial requires anticipating how one's actions will be interpreted by others. This requires a mentalistic theory-of-mind: the capacity to interpret behavior as being driven by beliefs, desires and intentions (Dennett, 1989). The same choice made in a different context

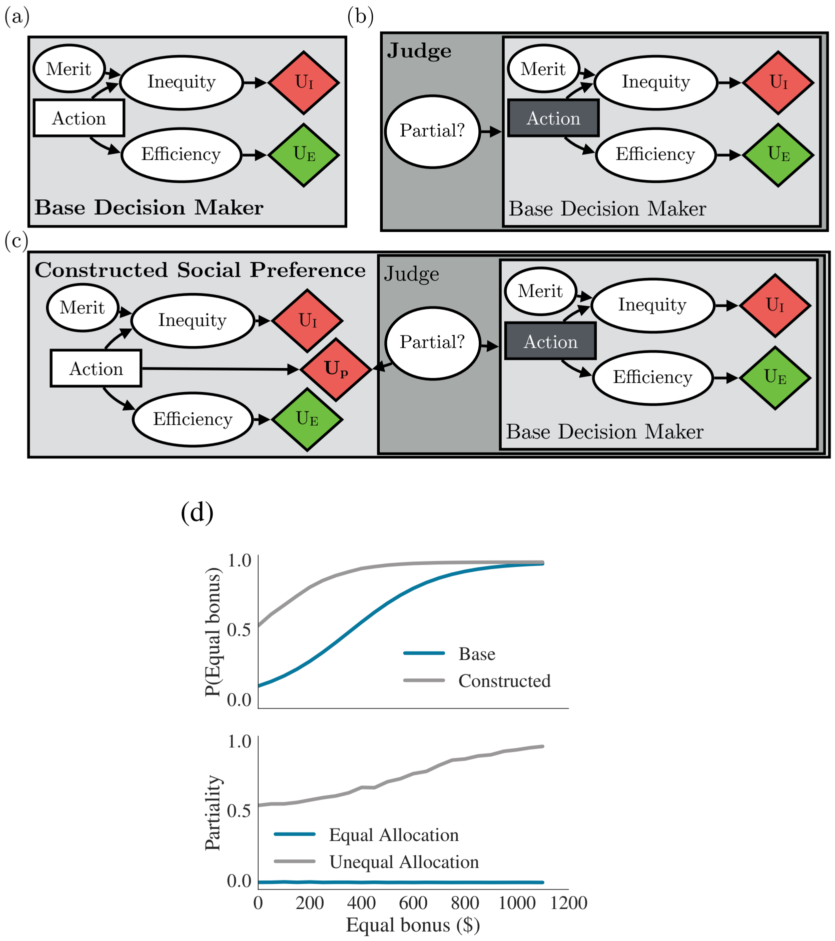


Figure 6-1: An influence diagram (ID) is a directed acyclic graph over three types of nodes: state nodes (circles), decision nodes (rectangles), and utility nodes (diamonds). Directed edges between nodes determine causal dependencies. State and utility nodes take values that depend on the values of their parent nodes. The total utility to the decision maker is the sum over the utility nodes. Green and red utility nodes correspond to rewards and costs respectively. The value of decision nodes is freely chosen by the decision making agent according to equation (6.4). (a) ID of the *Base Decision Maker*. Merit corresponds to $\boldsymbol{\gamma}$ and the Inequity and Efficiency nodes corresponds to the first and second components of equation (6.3) (b) ID of the *Judge* which infers whether a base decision maker was partial given an observation of her action, $P(\texttt{partial}|a)$. (c) The *Constructed Social Preference* recursively builds on the *Base Decision Maker* adding an aversion to appearing partial (U_P). (d) Simulated results when the decision maker can allocate $1,000 to one agent and $100 to another or the value on the x-axis to both agents when both agents are equally meritorious. The *Constructed Social Preference* is more likely to select the wasteful equal option to avoid an attribution of partiality.

or from a different set of alternatives might be evaluated differently as it will carry different information about the underlying goals and desires that drove the choice. For instance, if a decision maker can choose to give his colleague either $100 or $1,000 and chooses to give him $1,000 we might infer that he likes his colleague. However if his choices were to give either $1,000 or $2,000, giving $1,000 signals a dislikes for his colleague. Thus the same action requires a different interpretation depending on the unchosen option. Furthermore, the capacity for theory-of-mind can affect distributional preferences: previous work found that children with a more developed theory-of-mind were more likely to give fair offers in the ultimatum game (Takagishi, Kameshima, Schug, Koizumi, & Yamagishi, 2010).

In this work, we propose that preferences over the beliefs others will form are constructed by turning theory-of-mind inward, anticipating the evaluations others will make about the actions one might take. With the knowledge of how one's actions will be judged before deciding, a decision maker can calibrate her actions to send the right signals (Baumeister, 1982; Bénabou & Tirole, 2011). We note that we do not believe agents to be necessarily intentionally signaling impartiality to others. Instead agents may strive to maintain a desired image of themselves from an objective viewpoint or "self-signal" (Nagel, 1986; Bodner & Prelec, 2003; Bénabou & Tirole, 2011).

In this paper we develop a computational framework for capturing the above intuitions. We use influence diagrams as a structural representation of a rational actor and Bayesian inference over influence diagrams to enable theory-of-mind inferences about whether an action will be perceived as partial. While the framework we will present is a general way of constructing preferences from the anticipated judgments of others, we focus specifically on constructing distributional preferences with the desire to be perceived as impartial (Shaw, 2013; Shaw & Olson, 2014; Dungan, Waytz, & Young, 2014; DeScioli, 2016). We first present a mathematical model that integrates preferences for efficient and equitable outcomes with an aversion to appear partial. We then test our model empirically in two parameterized allocation games with many conditions that allow us to test some of the fine-grained predictions of the model. Finally, we conclude by sketching how our model can be extended to capture other social desires constructed from a decision maker's preference to appear positively in the minds of others.

6.2 Computational Analysis

In this work we aim to model both the way participants act in resource allocation games as well the judgments they make about the resource allocations of others. We start from the simpler preferences for efficiency and equity which are based on outcomes and build towards constructing a social preferences for impartiality which are implicitly intentional.

We define a resource allocation game as follows. Let $\mathcal{A}$ be the set of actions available to the decision maker. For each action $a \in \mathcal{A}$ there is a probabilistic transition function $P(R|a)$ which maps an action to a vector of rewards R where each $r_i \in R$ is the amount of reward given to agent i. In a resource allocation game, the decision maker picks an action (a) such that the expected reward to the other agents (R) achieves the desires of the decision maker.

We now define the desires of the *Base Decision Maker* as components of a utility function. These desires will determine how *Base Decision Maker* distributes resources. We consider two base desires. The first is a relative preference over the rewards received by specific agents. To realize this preference, we include the reward received by each of the other agents as weighted components of the decision maker's own utility. Depending on the value of these weights, an agent might impartially value others or might be partial towards certain individuals. Formally, let $\alpha_i \in \boldsymbol{\alpha}$ be the weight that the decision maker places on the reward given to agent i. When $\alpha_i > 0$, the decision maker gains utility proportional to the reward received by i, when $\alpha_i < 0$ the decision maker loses utility proportional to the reward received by i and when $\alpha = 0$ the decision maker is indifferent to the reward received by i. By expressing different α over different agents the decision maker can express partiality (or aversion) towards specific agents. Including the rewards received by all others as positive elements ($\alpha > 0$) in the decision maker's own utility creates a preference for Pareto efficient allocations, a form of efficiency where the reward distributed cannot be increased by taking other actions without making one of the receiving agents worse off.

The second base desire implements a form of proportional equity, the idea that those who contribute more to a joint endeavor should reap a larger share of the rewards or "just-desserts". A well studied way to capture proportional equity quantitatively is to constrain the relative reward (r_i) given to each agent to be proportional to their relative effort or merit

(γ_i) (Adams, 1965):

$$\frac{r_1}{\gamma_1} = \frac{r_2}{\gamma_2} = \ldots = \frac{r_N}{\gamma_N} \tag{6.1}$$

We transform these constraints into a measurement of inequity:

$$I(R, \boldsymbol{\gamma}) = \sum_{i \in N} \sum_{\substack{j \in N \\ j > i}} |\gamma_j r_i - \gamma_i r_j| \tag{6.2}$$

With a notion of efficiency and equity in place, we can define the allocation preferences for the *Base Decision Maker*. The expected utility (EU) to the decision maker of choosing a is:

$$\mathrm{EU}_{\texttt{base}}[a] = -\alpha_{IA}\mathrm{E}_a[I(R, \boldsymbol{\gamma})] + \sum_{i \in N} \alpha_i \mathrm{E}_a[r_i] \tag{6.3}$$

where $\mathrm{E}_a[I(R, \boldsymbol{\gamma})]$ is the expected amount of inequity created by action a and $\alpha_{IA} \in \boldsymbol{\alpha}$ is the weight the decision maker places on inequity aversion. $\mathrm{E}_a[r_i] = \sum_{r_i} r_i P(r_i|a)$ is the expected reward for i when the decision maker takes action a. Decision making follows probabilistically by sampling from the soft-max of expected utility:

$$P(a|\boldsymbol{\alpha}) \propto \exp(\beta * \mathrm{EU}[a]) \tag{6.4}$$

with higher values of β leading to a higher probability of selecting the action with the highest expected utility.

Influence diagrams are a natural choice for structurally representing this model since they can flexibly capture decision problems with multiple factors and recursive sources of value. Furthermore, they can be used to reason about the latent mental states of a decision maker from just a sparse and noisy observation of behavior (Jern & Kemp, 2015; Kleiman-Weiner et al., 2015). The utility of the *Base Decision Maker* which is defined in equation (6.3) can be expressed graphically as the influence diagram shown in Figure 6-1a. The first term of equation (6.3) corresponds to the U_I node and the second term corresponds to the U_E node.

We now consider a *Judge* who makes inferences and judgments about the underlying preferences of the *Base Decision Maker* following an observation of behavior. Specifically, in the *Base Decision Maker* the $\boldsymbol{\alpha}$ encode the preferences of the agent and so for the *Judge* these $\boldsymbol{\alpha}$ become the target of inference. For our purposes, the *Judge* is interested in the extent that the *Base Decision Maker* is partial to one or more agents. The *Judge*'s prior is that the *Base Decision Maker* is `partial` (a binary variable) with probability 0.5. If `partial`, one of the $\alpha_i = \alpha_{\texttt{partial}}$ (i chosen uniformly at random) and the other $\alpha_{-i} = -\alpha_{\texttt{partial}}$. Otherwise, if the agent is not `partial`, all $\alpha_{1\ldots N} = 1$. The *Judge* also has some prior uncertainty on the degree that the *Base Decision Maker* cares about inequity so $\alpha_{IA} \sim \text{Exponential}(\lambda)$. With these priors over the types of preferences a *Base Decision Maker* might have, a *Judge* can use Bayesian inference to compute the extent that an agent was partial based on just a single observed allocation:

$$P(\texttt{partial}, \boldsymbol{\alpha}|a) \propto P(a|\boldsymbol{\alpha})P(\boldsymbol{\alpha}|\texttt{partial})P(\texttt{partial}) \tag{6.5}$$

where $P(a|\boldsymbol{\alpha})$ is the model of action shown in equation (6.4) and the $\boldsymbol{\alpha}$ are then marginalized out to obtain a posterior on $P(\texttt{partial}|a)$. Figure 6-1b shows how the judge does inference over the parameters of the influence diagram representing the *Base Decision Maker*.

A *Constructed Social Preference* inherits from and recursively builds upon both the *Base Decision Maker* and the *Judge*. In particular, the *Constructed Social Preference* has an additional preference to appear impartial. Since this is a preference over the beliefs others will form as a result of her decision, the preference to appear impartial is a preference over the posterior $P(\texttt{partial}|a)$. The *Constructed Social Preference* integrates these belief based preferences with the preferences for equity and efficiency of the *Base Decision Maker*:

$$\text{EU}_{\texttt{constructed}}[a] = \text{EU}_{\texttt{base}}[a] - \alpha_{PA}P(\texttt{partial}|a) \tag{6.6}$$

where α_{PA} is the extent that the *Constructed Social Preference* cares about whether other agents view her as impartial or not. This equation and the influence diagram in Figure 6-1c show how the *Constructed Social Preference* is built on top of the *Judge* and *Base Decision*

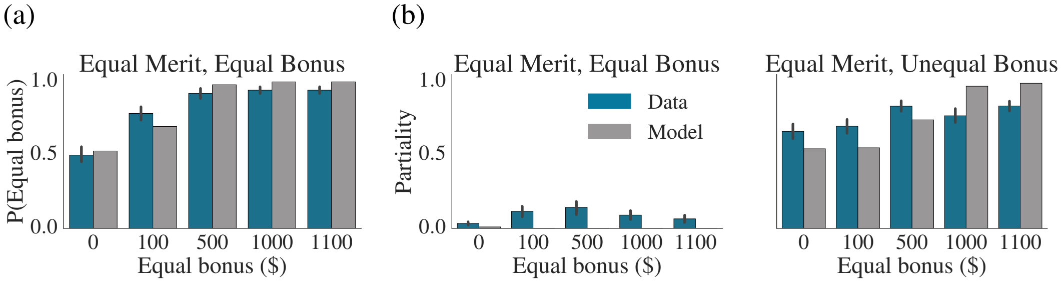


Figure 6-2: Empirical results and model predictions of (a) choices and (b) judgments of partiality for the trials in experiment 1 where both of the agents were equally meritorious. Trials with no gray bar indicate the model predicted near 0. Error bars are the standard error of the mean.

Maker.

The *Constructed Social Preference* goes beyond preferences over outcomes like those in the *Base Decision Maker*. Instead, it anticipates the inferences other agents will make about its actions and optimizes its actions so that others have desirable beliefs. Figure 6-1d shows a simulated example where a decision maker had to choose between allocating either $1,000 to one agent and $100 to another equally meritorious agent or giving a smaller but equal value to both. The *Constructed Social Preference* is more likely to select the equal option since it implies lower partiality even though both the *Base Decision Maker* and the *Constructed Social Preference* care equally about avoiding inequity.

In order to compare the model with human participants, we used maximum-likelihood estimation to optimize the free parameters to human judgments. The five parameters used for all simulations were: $\beta = 0.003$, $\alpha_{\texttt{partial}} = 6$, $\lambda = 0.7$, $\alpha_{PA} = 1350$. If agent i was more meritorious than agent j then $\frac{\gamma_i}{\gamma_j} = 4$. Importantly, the parameters used to model the partiality data were constrained to be the same as those used to model participants' decisions.

6.3 Experiments and Results

We test the predictions of this model in two parametric behavioral experiments that measure participants' decisions in a hypothetical resource allocation game as well as judgments about the partiality of another agent who made an allocation. Both experiments were run

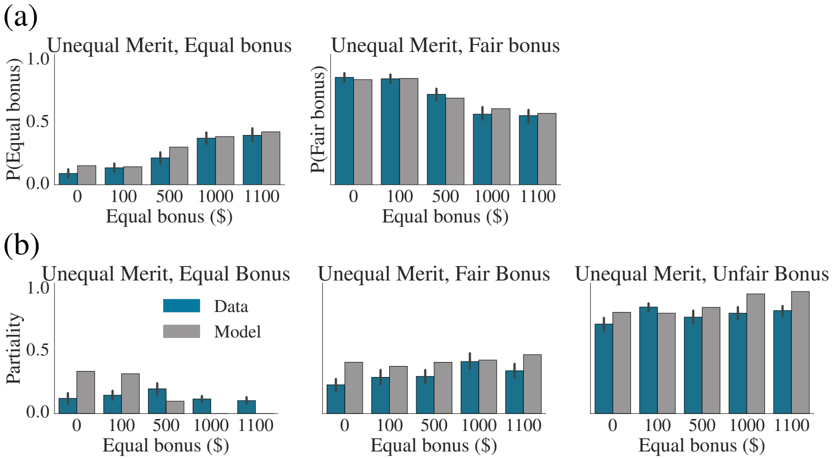


Figure 6-3: Empirical results and model predictions of (a) choices and (b) judgments of partiality for the trials in experiment 1 where one of the agents was more meritorious than the other. Trials with no gray bar indicate the model predicted near 0. A "fair bonus" was when the decision maker gave the large bonus to the agent with more merit. An "unfair bonus" was when the decision maker gave the large bonus to the agent with less merit. Error bars are the standard error of the mean.

on Amazon Mechanical Turk. For each condition we compare the average responses with the predictions of the model.

6.3.1 Experiment 1: Proportionality and Impartiality

In experiment 1 we investigate how equity and merit affect choices in an allocation game. We presented two groups of participants with the following vignette which describes an allocation game that took place in an everyday office setting:

> *Alex and Josh are both employees at a large company. Their coworker Max has been asked to decide how to assign bonuses to Alex and Josh. Due to company policy, Max can either: give $1,000 to one employee and $100 to the other or give* ***[$0 / $100 / $500 / $1000 / $1,100]*** *to both. Alex and Josh currently make the same amount each year, do the same job,* ***[and have received identical work evaluations / but Alex has received a better work evaluation]****.*
>
> Participant group 1: *What would you do? (Give Alex the $1,000 bonus and Josh the $100 bonus / Give Josh the $1,000 bonus and Alex the $100 bonus / Give them both a bonus of [$0 / $100 / $500 / $1000 / $1,100])*

> Participant group 2: *Max decides to* ***[give Alex the $1,000 bonus and Josh the $100 bonus / give Josh the $1,000 bonus and Alex the $100 bonus / give them both a bonus of ($0 / $100 / $500 / $1000 / $1,100)].*** *Who do you think Max likes better? (Definitely Alex = -1, Equal = 0, Definitely Josh = 1)*

The bold text shows the different variants of the vignettes. On different trials the value of the equal option varied between $0 and $1,100. On some trials both employees received equal work evaluations and on some trials one employee received a better work evaluation. The names of the employees changed on each trial but were always a high frequency male name.

We first report the results for when both employees were equally meritorious (Figure 6-2). We found high rates of inequity aversion that led to highly wasteful bonus allocations (Choices: N = 89; Judgments: N = 104). When the equal sized bonus was $0, almost 50% of participants chose to allocate nothing, wasting a total of $1,100 ($1,000 + $100) rather than allocating unequal bonuses. When the bonus was $100, over 75% of participants wasted the $1,000 bonus in favor of two equal $100 bonuses. These allocations were highly wasteful and were Pareto dominated since the unequal allocation would have made at least one of the employees better off without making the other employee worse off.

The partiality judgments made by a second set of participants is consistent with the idea that the aversion to creating unequal outcomes stems in part from a desire to appear impartial. We transformed judgments of liking into a partiality index by measuring absolute difference from 0. Even when the alternative equal allocation required wasting the entire bonus, a person who allocated the large but unequal bonus was judged as highly partial (towards the person who received the higher bonus). Our computational model corroborates this interpretation and captures both participants' judgments of partiality and then uses those judgments to explain the strong aversion to an unequal outcome. The full model closely follows the pattern of decision making.

We now turn to the trials where one of the two employees received a better evaluation at work than the other and was thus more meritorious (Choices: N = 89; Judgments: N = 104). Figure 6-3 shows that this difference was sufficient to drive participant choices away from the wasteful equal bonus towards giving the large but unequal bonus to the employee who was more meritorious. This shift is consistent with equity (the more deserving employee got a greater share of the rewards). However, this also resulted in a novel type of wasteful decision making: the option to allocate \$1,000 or more to both employees was forgone over 70% of the time by the Pareto dominated unequal option that maintains equity based on merit.

Surprisingly, participants attributed the lowest partiality to employees who selected the equal bonus even though one of the receiving employees was more deserving than the other. This points to a possible difficulty in achieving equitable distributions. Even when some agents might be more deserving than others, inferences of partiality are still readily made when observing an unequal distribution. Here equity and impartiality work against each other. Since the equal bonus led to a lower attribution of partially, as the size of the equal bonus grows, the model slowly shifts to the efficient equal bonus.

6.3.2 Experiment 2: Procedural Fairness and Impartiality

In a second experiment we repeated the equal merit condition of experiment 1 but also included the possibility that the employee making the decision could flip a fair coin to decide who gets \$1,000 and who gets \$100 (Choices: N = 54; Judgments: N = 158). Besides the addition of this coin the vignette was identical to the vignette in experiment 1. This is a key test of the impartiality hypothesis since when the size of the equal bonus is low, an inequitable but efficient allocation can be given *without* signaling partiality towards either of the employees by flipping a coin (Shaw & Olson, 2014; Choshen-Hillel et al., 2015).

Consistent with the model predictions shown in Figure 6-4, participants did not judge employees who flipped the coin to be partial towards either of the employees. When the value of the equal bonus was low ($\leq$ \$100) participants no longer wasted resources like they did in experiment 1. Instead they flipped the coin in order to allocate the full bonus without signaling partiality.

Combining the two experiments, we quantify the overall model performance across all of the conditions in the two experiments. Figure 6-5 shows the quantitative correlation

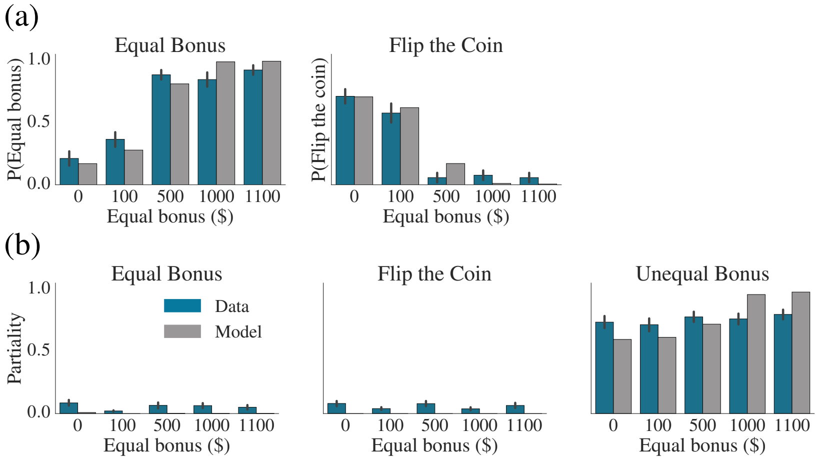


Figure 6-4: Empirical results and model predictions of (a) choices and (b) judgments of partiality for experiment 2 which introduced the option to flip a fair coin to decide the allocation of the unequal bonuses. Trials with no gray bar indicate the model predicted near 0. Error bars are the standard error of the mean.

of the model predictions with the average judgments of participants. Overall, participant judgments and decisions were highly correlated ($R^2 = 0.94$) with the model predictions. This suggests that the model is capturing some of the fine grained structure of how people attribute both partiality and use it to make allocations of welfare.

Finally, we compare the full model presented here against a lesioned model that includes inequity aversion but does not reason about partiality and hence corresponds to the *Base Decision Maker* (i.e., $\alpha_{PA} = 0$). The parameters in the lesioned model were directly fit to the choice data and were not constrained to fit the judgments. This model fit the data less well than the full model ($R^2 = 0.82$). However, this lesioned model has less parameters than the full model. To test for the possibility that the full model is overfitting the data we performed cross-validation using randomly chosen subsets of half the data to fit the free parameters and then tested against the held-out half. The held-out cross-validation correlation between the model and participants was $R^2 = 0.93$ which suggests that the full model is robust and is not overfitting. In contrast, the lesioned model performed much worse ($R^2 = 0.74$) under cross-validation. When the full model was applied only to the choice data it captured nearly all of the variance ($R^2 = 0.97$) and was still robust when evaluated on only held-out trials ($R^2 = 0.96$).

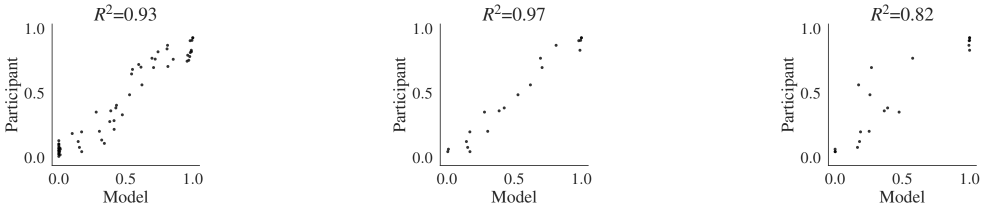


Figure 6-5: Quantification of model performance. Each point represents the model prediction and participant judgment for a single condition. For better fitting models the points will lie close to the $y = x$ diagonal. (left) The full model compared including both decision and judgment data. (middle) The full model compared only on the decision data. (right) Lesioned model that did not include partiality compared only on the decision data.

6.4 Discussion

We introduced a new computational model for constructing preferences by modeling rational agents which care about what others will infer about them from their actions. In this model, the machinery of theory-of-mind is turned inward to simulate how an action will likely be perceived or judged by others. Agents then use the perceptions and judgments they anticipate others will form to construct rich preferences over socially desirable traits such as impartiality. We tested key components of the model in two behavioral experiments that were designed to contain conflict between efficiency, equity and partiality and measured both participants' hypothetical resource allocations and the judgments they made about the partiality of others who had acted. The predictions of the model were closely correlated with both allocation decisions as well as partiality judgments. Finally, we note the best fit parameters had a high value for α_{PA} which suggests that partiality aversion was playing an important role in the model fit for predicting choices. A lesioned model that did not contain this parameter failed to predict participants' judgments in both experiments.

We now briefly describe qualitatively some of the other predictions this model can make without any structural extension. Our model predicts that when the decision maker and one of the agents have a previous relationship (such as old friends or a reciprocal relationship in a different context) there will be a greater probability of inferring partiality since this previous relationships will manifest itself on the prior over `partial`. With a greater probability of others inferring partiality a decision maker will be even less likely to give their friend a larger reward than another person. This reasoning might explain why nepotism and cronyism is judged as unfair and avoided (Dungan et al., 2014). Other procedural tools such as the delegation of the decision to a third party may also be important to avoid the attribution of partiality. Under the model we have presented, if an attribution of partiality can be made less likely, the decision maker might be more likely to participate in nepotism and favoritism.

In future work we would like to investigate how other forms of social preferences can be constructed by placing preferences over anticipated judgments. For instance, people might desire to appear as trustworthy and generous or avoid appearing selfish or envious. Ultimately we suspect that an agent who carefully manipulates their image so that all others think she is a great person – will end up behaving quite similar to a person who is truly good. However, her behavior will be less robust – when she suspects her actions are unobserved or can only be interpreted ambiguously, the constructed social preferences disappears along with the altruistic or fair behavior (Dana, Weber, & Kuang, 2007). By constructing social preferences such as impartiality, a key component of fairness, from the anticipated judgments of others, we quantitatively predict the fine-grained structure of both participants' decisions concerning the allocation of resources and participants' judgments about those who make distribution decisions. Our model makes clear that the power of theory-of-mind is not necessarily limited to understanding the beliefs and desires of other intentional agents. It can also be pointed inward to strategically shape beliefs and desires in others.

Chapter 7

Inference of Intention in the Computation of Moral Judgment

7.1 Introduction

Our actions often have multiple effects, whether it's creating a small amount of pollution in order to pick up groceries or making trade-offs between civilian deaths and military objectives during a war. Did the general try to achieve the military objective even at the cost of civilian lives or did his plan use civilian deaths in order to demoralize the enemy? The ability to distinguish between the effects an agent intended versus those that were side-effects are critical in general for social cognition and in particular for assigning responsibility and assessing moral permissibility. Our goal here is to understand these processes in computational terms.

Reasoning about the intentions of other agents relies on theory of mind, the capacity to infer an agent's underlying mental states such as beliefs and desires from her actions. Recently, a lot of progress in computational modeling of theory of mind has been made by formalizing lay intuitions that other agents act as rational actors who maximize expected utility subject to their beliefs. A Bayesian observer can then invert the agent's planning process and reason about the likelihood of certain beliefs and desires given the agent's actions (C. L. Baker et al., 2009, 2017; Jern & Kemp, 2015).

Although most computational accounts of theory of mind have focused on desires and beliefs, intentions are a third mental state thought to be particularly useful. Intentions can be thought of as plans of action that an agent commits to, chosen in order to bring about its desires given its beliefs about the causal structure of the world (M. Bratman, 1987; Malle & Knobe, 1997). Using an example from M. Bratman, 1987, imagine a person with the desires to have a milkshake but also the desire to go on a diet. If the desire to go on a diet is sufficiently high, the person will not buy the milkshake without any irrationality on the part of the decision maker. In contrast, the person cannot intend to both go on a diet and have the milkshake. Thus while desires can be in conflict an intention must have a sense of coherence and commitment. Further, Malle and Knobe argue that three key criteria differentiate intentions from desires: “First, intentions are directed at the intenderâĂŹs own action, whereas desires can be directed at anything. Second, intentions are based on some amount of reasoning whereas desires are typically the input to such reasoning. Third, intentions come with a characteristic commitment to perform the intended action, whereas desires do not” (Malle & Knobe, 2001). Children understand these mental states early in development and begin to understand the subtle between desires and intention as early as 5 years old (Schult, 2002).

It is hypothesized that the ability to reason about and with the intentions of others is one the key factors that enables the sophistication of human social behavior (Tomasello et al., 2005; Tomasello, 2014). They are also an important input into the evaluation of moral permissibility such as the doctrine of double effect’s requirement against intending harm (Mikhail, 2007; Cushman, 2013; Waldmann, Nagel, & Wiegmann, 2012; Crockett, 2013; Greene, 2014). Specifically the doctrine of double effect (DDE) states among other an act is morally permissible if:

1. the action itself is not morally impermissible,
2. the good outcomes but not the bad outcomes are intended,
3. there is no way to produce the good outcomes without also producing the bad outcomes,
4. the bad outcomes are not disproportionate to the good outcomes (Mikhail, 2007, 2011).

Beyond normative theories, intention is also a key factor in determining legal culpability. People are held responsible and punished for just having an intention to cause harm (*mens rea*) even if no harm is actually caused (L. Young, Cushman, Hauser, & Saxe, 2007; Cushman, 2008). Again, there is a key distinction between desire and intention. One might have the desire to harm ones enemy (or benefit oneself even at the expense of someone else), but until those desires have actually generated an intention they are not punished.

While the DDE gives intentions a prominent role in moral judgment, the relationship between intentions and outcomes is complicated by the fact that it is possible to do the right thing for the wrong reasons (Scanlon, 2009). Scanlon, 2009 imagines cases such as saving a person in need but only to receive the fame or a case where an agent saves the mother of his political opponent only so that his political opponent doesn't receive his mother's large inheritance to spend on the campaign. Scanlon argues that even though the intentions aren't correct the actions themselves are still permissible.

Here we investigate a novel computational representation for reasoning about other people's intentions based on counterfactual contrasts defined over influence diagrams. This model can distinguish between intended outcomes and unintended side effects as well as represent the future-oriented aspect of intentions as plans (M. Bratman, 1987). We use this model of intention inference as an input into a computational model of intuitive moral permissibility judgments based on the DDE and show this model explains both well-studied and novel moral dilemmas. We first motivate our approach with examples to build intuitions. We then describe a computational model that captures these intuitions. Finally, we show how this model can explain human moral judgments across a wide range of moral dilemma.

Throughout this work we will motivate and test our model on variations on the well-known "trolley problems" (Thomson, 1985). Although lacking in everyday realism and heavily studied in moral psychology, our aim here is not to explain a novel phenomena in moral judgment, instead our aim is synthesize known aspect of moral psychology into a unified model of moral judgment. For this purpose, "trolley problems" are highly suitable. They are familiar to both readers and participants with clear causal structures, events, and can be easily parameterized.

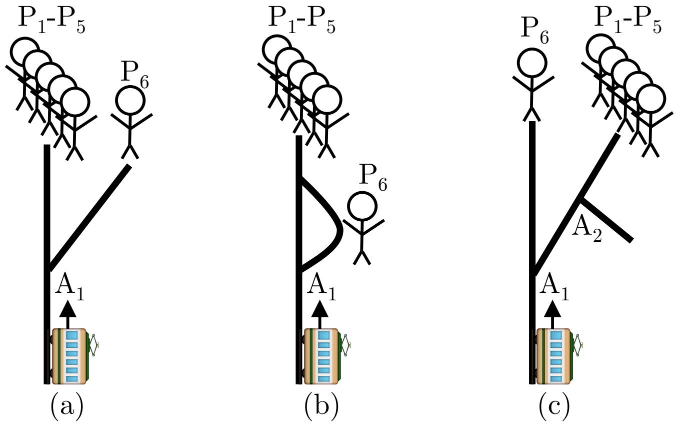


Figure 7-1: Schematic representation of trolley track geometries: (a) side track, (b) loop track and (c) side-side track.

7.1.1 Side track and loop track

The canonical examples for the role of intention in moral permissibility judgments are the *side track* and *loop track* (Thomson, 1985). The *side track*, shown in Figure 7-1a is a scenario where an out-of-control trolley is heading towards five people. An agent is standing near a switch (A_1) which will turn the trolley from the main track with five people on it (P_1-P_5) to a side track with one person (P_6). The *loop track*, shown in Figure 7-1b has a loop instead of a split such that the trolley will continue on and hit the five unless it hits the man on the looping track which would cause the train to stop. Consider that in each of the situations, the agent throws the switch.

Empirically, throwing the switch in the *loop track* is judged less morally permissible than the *side track* (Mikhail, 2007). Explanations of this finding usually draw on the agent's intention. In the *side track*, the agent neither intends the hitting nor killing of the man on the side track while in the *loop track* the agent does intend for the trolley to hit the man on the loop but not his death. Indeed, Sinnott-Armstrong, Mallon, Mccoy, & Hull, 2008 found that people judge the decision maker as both more responsible and intentional in the *loop track* case than the *side track*.

7.1.2 Side-side track

Following M. Bratman (1987), future-oriented planning is an important aspect of intention. To probe this aspect of intention in permissibility we developed a novel track geometry which requires inference over the full plan rather than just a single action. As shown in Figure 7-1c, the *side-side track* scenario is similar to the *side track* except that the side track has an additional side track with its own switch (A_2). Consider a situation in which there is one person on the main track, five people on the side track and no one on the side-side track. If the trolley is going down the side track, unless the agent throws the second switch directing the trolley down the side-side track, the trolley will continue and hit the people on the side track.

We hypothesize that throwing the first switch is intuitively morally permissible. How can this be explained even though the trolley is now heading towards the five people? Since intentions are forward-directed, they include the agent's intention to throw the second switch, saving all the lives. Only if the agent doesn't intend to throw the second switch does the action become impermissible. Thus this notion of intention must clearly influence moral permissibility, without knowing the agents future intentions, we cannot assess the outcomes. This case motivates the central role of planning in our computational model. It is insufficient to consider intentions as merely directed towards the effects of a single action but rather the effects of the entire plan need to be taken into consideration.

7.1.3 Joint inferences: norms, desires and intentions

Inferences about the intentions of an agent are often intertwined with inferences about the agent's desires and the social norms to which those desires conform. We contrast a *side track* dilemma that has one anonymous person on the main track and two anonymous people on the side track with a dilemma we call *brother track* where the agent's brother is on the main track and there are two anonymous people on the side track.

In the first dilemma, throwing the switch is likely to be judged as highly impermissible while in the *brother track* case it is intuitively more permissible to save one's brother. When the agent throws the switch in the first case, participants might infer that the agent intended to kill the two people. In *brother track*, participants infer that the agent is following a norm to value loved ones more and doesn't intend to kill the two people on the side track, instead the intention of their action was to save their brother. In another variant of *brother track*, the brother is on the side-track and two anonymous people are on the main track. If the agent throws the switch, we may infer that the agent followed a "hyper-egalitarian" norm that all lives should be valued equally, or... she might not value all lives equally, she just intended to kill her brother! To infer the intended consequences and judge moral permissibility thus requires jointly inferring the agent's desires and the norms that guided their actions.

This problem of joint inference also arises when assessing the moral permissibility of decisions made under uncertainty. Consider the *side-side track* case but the second switch has a 90% probability of failure. If the decision maker throws the switch they save 1 person with certainty but will kill 5 people with 90% probability. Thus this decision will lead to 3.5 lives lost in expectation. While this could mean the decision maker intends to kill those people, another interpretation is that the decision maker is risk seeking in this situation, preferring a small chance of saving everyone than letting one person die with certainty. If the attitude of the decision maker towards risk is taken into account this can change the inferred intention and hence also change whether or not their action is seen as morally permissible.

These kinds of complications also arise when agents have to make taboo trade-offs. Consider a problem, *equipment track* with the same structure as *side track* but the decision maker is the owner of the train company and in addition to the people on the tracks, there is piece of equipment on one of the tracks that is highly valuable to the train company. If the owner throws the switch, saving the five people and the piece of train equipment, but killing one person on the side track was it done for the right reasons? Again, this judgment requires joint inference over the value the owner places on the conversion rate between the value of the taboo good (monetary value) and the lives saved and killed. Furthermore, imagine positions of the people are switched, with only one person and the equipment on the main track and five people on the side track. If the owner throws the switch, out intuition isn't that the owner intends to kill the 5 people nor that he intends to save the single person. Instead, we infer that he simply intends to save the equipment.

These three cases, *brother track*, *side-side track*, and *equipment track* will be modeled in detail across a wide range of parametric variations that aim to push around these joint inferences. We then test the predictions of this model empirically in behavioral experiments.

7.2 Computational Framework

Our computational approach has two parts. The first is a computational account of intention inference and the second uses this account to model permissibility judgments. The model is presented to capture the real-world richness of intentional planning and has greater generality than is needed for our examples.

Our representation of intentions is based on influence diagrams (ID) (Shachter, 1986). Influence diagrams are similar to Bayes nets and were used previously to capture reasoning about what other agents know and want during decision-making (Jern & Kemp, 2015). Solving an ID yields an optimal policy (σ^*): the actions the decision-making agent needs to take to maximize her expected utility. We show how the ID and the policy can be used together to compute foreseen outcomes: the most likely outcome of the agent's policy. Using a counterfactual criterion, we refine the foreseen outcomes into a subset of outcomes that are intended.

Overall, we aim to capture that intentions: (1) are partial plans with means-ends correspondence, (2) predict the expected effects of actions, (3) can distinguish between outcomes that the agent is committed to bring about and those that are side-effects, (4) are future-oriented, (5) give reasons for action and are hence inputs to further practical reasoning such as moral permissibility (M. Bratman, 1987). Indeed, one practical reason for the centrality of intentions in folk psychology is that knowing an agent's intentions allows one to predict how the agent will behave and why.

We then show how an observer with uncertainty about the desires and norms of the agent can rationally update his beliefs about the agent by inverting the planning process using Bayes' rule, and finally, can use these inferences to make judgments about moral permissibility.

7.2.1 Rational action in influence diagrams

We first introduce influence diagrams which generalized the factored representations of Bayesian networks to decision problems. We follow the notation of Koller and Friedman (2009). An influence diagram $\texttt{ID}$ is a directed acyclic graph over three types of nodes: state nodes (depicted as circles, $\mathcal{X}$), decision nodes (depicted as rectangles, $\mathcal{D}$), and utility nodes (depicted as diamonds, $\mathcal{U}$). Directed edges between nodes determine causal dependencies. State and utility nodes take values that are a function of the structural equations and depend on the values of their parent nodes. In particular, utility nodes take a real number as a value which when summed together represent the total utility to the decision making agent.

Unlike state and utility nodes, the value of decision nodes are not determined by structural equations but are instead freely chosen by the decision making agent. Under the assumption of rational action, the decision making agent attempts to make these decisions such that their total utility is maximized. The full set of these decisions in a given problem is an agent's *policy* which map each decision node to a particular value and is represented by σ. Let σ^* be the policy that maximizes the expected total utility of the decision problem represented by $\texttt{ID}$:

$$\sigma^* = \arg\max_{\sigma} \mathrm{EU}[ID_\sigma]$$

where $\mathrm{EU}[\texttt{ID}_\sigma]$ is the agent's expected utility when following policy σ. In order to calculate the expected utility of a given policy, we define ζ as an *outcome*, the setting of each of the state, utility and decision nodes in *ID* to a value. For each node $Z \in \texttt{ID}$, ζ_Z is the value of node Z in outcome ζ. Thus the expected utility of policy σ can be calculated by averaging the total utility of an outcome $U(\zeta)$, weighted by the likelihood of that outcome under the policy $P(\zeta|\texttt{ID}_\sigma)$ for each possible outcome:

$$\mathrm{EU}[ID_\sigma] = \sum_{\zeta} P(\zeta|ID_\sigma)U(\zeta)$$

$$U(\zeta) = \sum_{V \in \mathcal{U}} \zeta_V$$

$$P(\zeta|ID_\sigma) = \prod_{X \in X} P(X|\mathrm{Pa}_X, \sigma)$$

where Pa_X are the parents of node X. Thus the ID representation concisely factors the agent's decision problem into individual states, decisions, sources of utility and the structural equations that define the dependence relations between them. Defaults are encoded by requiring any policy that changes the value of a decision node away from its default value to incur a small utility cost (not shown in figures).

See Figure 7-2a for an influence diagram representation of the *side track* dilemma shown in Figure 7-1a. The only decision in the policy is the choice to throw the switch A_1. This action determines whether the trolley goes down the left track (T_L) track or right track (T_R) which determines which people are hit and killed affecting the decision maker's utility. Specifically, if we assume that the utility of each person on the tracks to the decision making agent is 1, then the agent's policy under rational action is to throw the switch (since 5 is greater than 1). In Figure 7-2a this policy is shaded in gray. We will continue to build on this example as we build out our model of intention.

7.2.2 Foreseen and Intended Outcomes

The structure of the influence diagram allows us to reason explicitly about the possible outcomes that result from an agent's choices. We first describe foreseen outcomes that characterize the agent's expectations about the effect of her actions before refining the foreseen outcome into an intention, which excludes outcomes that are unintended.

Definition 2. The best foreseen outcome ζ_F is the outcome with the highest expected utility that can be foreseen by the agent acting under rational action:

$$\zeta_F = \arg\max_{\zeta} \mathrm{U}(\zeta) P(\zeta | \mathtt{ID}_{\sigma^*})$$

ζ_F captures all the consequences that the agent can optimistically foresee happening as a result of her policy but does not include backup plans or other types of conditional contingent plans. The decision to choose only a single foreseen state is motivated by efficient planning algorithms which plan only on the likeliest states and replan if necessary rather than exhaustively planning for every contingency (Platt, Tedrake, Kaelbling, & Lozano-Perez, 2010).

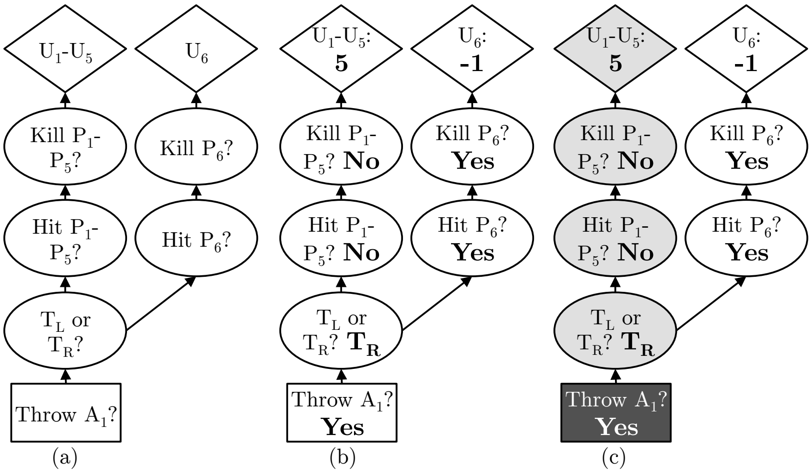


Figure 7-2: An influence diagram (ID) representation of intention. (a) The ID for the *side track* decision dilemma. (b) The foreseen outcomes F. Each node is set to the best value possible under the policy of throwing the switch (shown in bold). (c) The intention I is shaded in gray. Like the foreseen outcome, each node is set to its most likely value under the policy, however only the nodes shaded in gray and their values are intended by the agent.

In our working example of the side track, the foreseen effects of throwing the switch in *side track* are that the 5 people on the main track *will not* be hit by the trolley and live, generating 5 utility while the person on the side track *will* get hit by the trolley and die generating -1 utility for the decision-maker. This is shown in Figure 7-2b where each node is assigned to its foreseen value ζ_F (shown in bold). Thus the influence diagram allows for an intuitive causal representation of the effects of the agent's decision, each with its own utility nodes which can be reasoned about independently.

While foreseen outcomes optimistically describe the consequences of an action and are brought about "intentionally", not all foreseen consequences are intended by the decision maker (M. Bratman, 1987). Analogously in causal reasoning, not all of the factors which influence an outcome are judged by human participants to be causes of an observed outcome. This has led to the development of computational models of *actual causation* which try to model the commonsense notion of causality through counterfactual reasoning (Halpern & Pearl, 2005). This formalism has successfully captured aspects of empirical attribution of responsibility (Lagnado, Gerstenberg, & Zultan, 2013; Sloman, Fernbach, & Ewing, 2012; Chockler & Halpern, 2004). We propose that a similar model can distinguish an agent's intended outcomes from foreseen outcomes. Intended outcomes (as opposed to unintended outcomes), are those that are a *cause* of the agents particular plan. By phrasing this hypothesis in the language of causal reasoning we can use computational tools used to model reasoning about causal events such as counterfactuals to model reasoning about intentions.

These models determine whether an event X was a cause of an observed effect Z by checking whether or not Z counterfactually depends on X i.e., in a counterfactual world where X didn't occur the effect Z also doesn't also occur. This counterfactual operation captures the intuition that mere co-occurrence is insufficient for causation. If X and Z co-occur but X is not a cause of Z then in the counterfactual world where X doesn't occur, Z will still happen and hence X will not be considered a cause of Z.

Although this simple definition captures human intuitions about some causal situation it requires some extension to capture the robustness of human judgments, namely those about which the effect is overdetermined by multiple causes each of which is sufficient to bring about the effect e.g., if X or Y then Z. These cases are tricky for the simple counterfactual model because in the counterfactual world where X doesn't occur, Y still causes Z happen and hence by symmetry neither X nor Y is considered a cause of Z since Z did not depend on either variable in the counterfactual world. This fails to capture the intuition and data that *both* X and Y are causes of Z in cases of this structure. To robustly capture human judgments in these settings we employ the full counterfactual machinery developed by Halpern and Pearl which expand the space of counterfactual worlds which can be considered. For example in the case of overdetermination described, the model allows one to consider the world where neither X nor Y occur and so Z doesn't occur and hence both X and Y are predicted as causes. Thus these models can capture commonsense causal judgment through counterfactual reasoning which are thought to be an important input into other forms attribution and responsibility judgments (Lagnado et al., 2013; Sloman et al., 2012).

Halpern and Pearl counterfactuals were originally developed for reasoning about causal graphs and the structural equations that describe the relationships between the event nodes. Here we generalize Halpern and Pearl counterfactuals to reasoning about the causes of actions chosen by an agent by considering counterfactual contrasts over influence diagrams. Thus while counterfactuals over Bayesian networks answer a question about commonsense or *actual causation*, we propose that counterfactuals over influence diagrams can answer a question about *mental causation* which describe the intentions of the agent.

Specifically, intended outcomes are the subset of foreseen outcomes that the choice of the optimal policy (σ^*) counterfactually depends upon. The full formal machinery of Halpern and Pearl is crucial for reasoning about these intentions because in many cases an agents action is highly overdetermined i.e., there were many source of utility that when combined together were resulted in the specific plan of that agent. While overdetermination often requires carefully constructed cases in causal reasoning where causes happen simultaneously, overdetermination is omnipresent in intentional reasoning since there are usually many possible reasons for action. Intentional inference is to decide which of those reasons are intended by the agent. We generalize Halpern and Pearl (2005) to decision problems where the *effect* is actually the planned set of actions and the structural equations are supplemented by the action nodes of the influence diagram and rational planning. Our definition of an intention is as follows:

Definition 3. An intention I is a subset of nodes and their corresponding values such that the following conditions are satisfied:

1. Nodes in I take on values foreseen under σ^*.

2. Let $ID^{\setminus I}$ be a counterfactual influence diagram that is ID with the nodes in I fixed. I are intended if $\sigma^*_{\setminus I} \neq \sigma^*$, i.e., the optimal policy for $ID^{\setminus I}$ is different from the optimal policy for the original influence diagram ID.

3. The sets of nodes in I are a minimal subset, i.e., there are no smaller subsets of intended nodes, which when removed or fixed would also satisfy 2.

The intention I for the *side track* is shown in Figure 7-2c by the nodes and values highlighted in gray. The decision to throw the switch does not depend on the values for hitting and killing the person on the side track (P_6) and the loss of utility that resulted. Even if those nodes were removed from the influence diagram the agent would have still acted the same. Thus those nodes are side effects of the action. In contrast, if the nodes that correspond to the states and utility of the people on the main track were removed, the agent would not have thrown the switch. Since the policy with the nodes removed is not equal to the policy for the full *ID*, those nodes and their values are treated as intended. We only consider the fixation or removal of nodes. However, capturing other aspects of intention may require counterfactual perturbations to the utility values rather than fixation (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015).

Our representation of intentions as counterfactuals over influence diagrams satisfies the five aspects of intentions we aimed to capture: (1) *I* is a partial plan than contains future expected actions, (2) the outcomes in *I* are the expected result of the plan, (3) *I* distinguishes between intended outcomes the agent is committed to bring about and side effects, (4) *I* contains future-oriented policy information, (5) the nodes and values in *I* give the reasons for the action. By representing the intention as a subset of the influence diagram, we can capture the notion that intentions include *intend-to* commitments that commit the agent to future action as representing by intended action nodesx as well as *intend-that* commitments that commit the agent to make the world take on a specific state as represented by intended state nodes. These aspects of intention have been discussed in formal planning models and are thought to be important for collaborative planning (Grosz & Kraus, 1996). Thus *I* is a compressed representation of the actions an agent plans to make and their intended effects.

This is a useful representation in that the agent doesn't have to continuously plan after each decision and event. As long as things are going as expected (or "according to plan") the agent can continue to follow his policy. When things don't go as intended, the agent doesn't have to begin planning from scratch but can instead try to plan back to the intended state. We developed this model with the goal of representing intention inference in theory of mind and remain agnostic as to whether these computations capture the deliberate processes of a decision maker.

If *I* contains probabilistic contingencies, then we might prefer to use the word "hopes" rather than intends. It seems strange to say that an agent intends to win $5,000,000 when he buys a lottery ticket (if $5,000,000 is the highest payoff); it seems more reasonable to say that he hopes to win $5,000,000. Similarly, if a doctor performs an operation on a patient who has cancer that he believes has only a 30% chance of complete remission, it seems strange to say that he "intends" to cure the patient, although he certainly hopes to cure the patient by performing the operation. In addition, once we think in terms of "hopes" rather than "intends", it may make sense to consider not just the best outcome, but all reasonably good outcomes. For example, the agent who buys a lottery ticket might be happy to win any prize that gives over $10,000, and the doctor might also be happy if the patient gets a remission for 10 years.

The following example, which is due to Chisholm (1966) and discussed at length by Searle (1969), has been difficult for other notions of intention to deal with: Louis wants to kill his uncle and has a plan for doing so. On the way to his uncle's house in order to carry out his plan, he gets so agitated due to thinking about the plan that he loses control of his car, running over a pedestrian, who turns out to be his uncle. Although Louis wants to kill his uncle, we would not want to say that Louis intended to kill his uncle by running over the pedestrian, nor that he intended to run over the pedestrian at all. Given reasonable assumptions about Louis's beliefs (specifically, that the pedestrian was extremely unlikely to be his uncle), he clearly would have preferred not to run over the pedestrian than to run him over, so the outcome running over the pedestrian was not intended. Thus according to our definitions, even though he full-filled his desire, he did so unintentionally.

Example: Side track and loop track

As demonstrated before (see Figure 7-2c), the model correctly predicts that hitting and killing the man on the side track is unintended. In contrast, for the *loop track*, when the man on the loop is hit but not killed, the policy remains unchanged, so the model predicts that the killing of the man is unintended. However, the model predicts that hitting the man on the loop is intended since it is required to stop the trolley from hitting the 5 on the main track. Thus due to this difference in causal structure, the agent in *loop track* intends to hit but not kill the man on the loop. Indeed throwing the switch in *loop track* is found to be less permissible than throwing the switch in *side track*. Given that the number of lives affected is the same in both conditions suggests that the intention to harm in the loop track case could account for this difference as has been suggested in the literature (Mikhail, 2007).

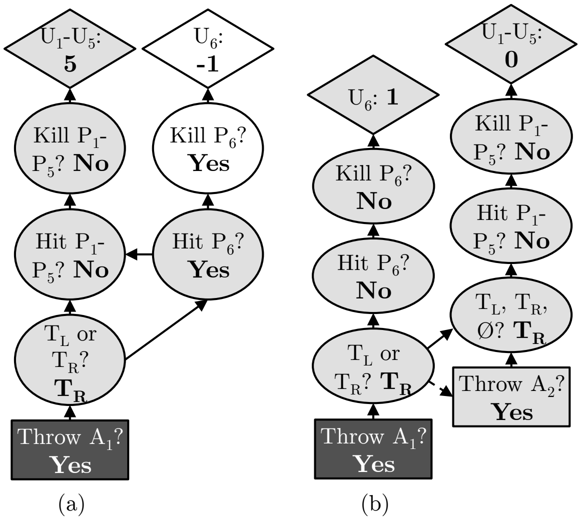


Figure 7-3: Influence diagram for the (a) *loop track* and (b) *side-side track* with the intention shaded in gray and the action in black.

Example: Side-side track

In the *side-side track*, the model predicts that if the agent throws the first switch, her intention is to also throw the second switch so that the trolley goes down the side-side track and kills nobody. The model further predicts that both saving the person on the main track and the 5 people on the side track are intended since in both cases they were counterfactually relevant to the policy: if the person on the main track wasn't there, the agent wouldn't throw the first switch. If the people on the side track weren't there, the agent wouldn't have thrown the second switch (since throwing switches has a small action cost associated with it).

The role of intention in evaluating permissibility is clear here even though it plays a different role than in the *loop track*. The number of lives affected can only be calculated under the agent's future-oriented plan. Thus the intention captures a key aspect of the permissibility by requiring an inference over future actions rather than through understanding which effects are intended and which are side-effects.

7.2.3 Intention inference through inverse planning

Inference of intention through the ID requires knowledge of the agent's desires and beliefs. However, observers often only know these desires and beliefs with uncertainty such as in the *brother track* examples in the introduction. The structure of these priors gives the observer an expressive theory of mind, capable of representing agents with both good and evil desires or adherence to different norms. The observer's beliefs about the agent's desires are modeled by introducing uncertainty over the parameterization of the utility nodes. This uncertainty induces a probability measure over IDs (shown in Figure 7-4) and since each ID has an intention under rational planning, it also induces a probability measure over intentions. Given observation of an agent's action(s) A, an observer can rationally update his belief about the agent's intentions I, desires D and norms N using Bayes rule:

$$P(I,D,N|A) = P(A|I)P(I|D,N)P(D,N)/P(A)$$

Since $P(A)$ cannot be analytically calculated we used rejection sampling to draw samples from $P(I,D,N|A)$. We first sample from the desire and norm distribution of the observer $P(D,N)$ which defines an influence diagram $ID^{D,N}$. Planning in this ID yields $P(I|D,N)$. If the intended action is the same as the observed action A we keep the sample which is a joint distribution over the intention, desires and norms. If the intended action is not A, the sample is discarded and the processes is repeated.

Structured priors for norms and risk preferences

In order to quantitatively predict observer's judgments of $P(I,D,N|A)$ we must specify the structure of $P(D,N)$, the distribution over how the agent values the lives of the people on the tracks. Let D_T be the utility to the decision maker of the n_T people on track T not being killed. Let $k_T = -1$ when the agent wants to kill the people on track T, if not, $k_T = 1$. k_T is negative for all T with probability α_b which means the agent wants to kill as many people as possible. Otherwise, $k_T = -1$ for each track independently with probability α_k. This allows for the decision maker to selectively want to kill a specific group of individuals (and

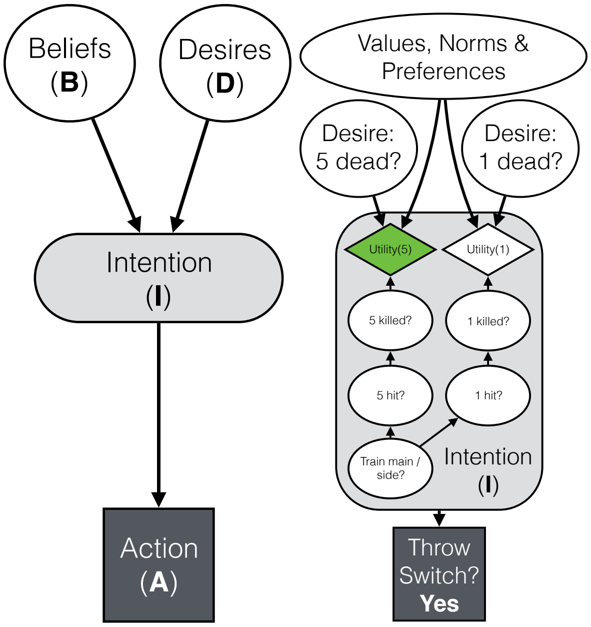


Figure 7-4: Joint inference of intentions, desires and norms in a belief, desire, intention (BDI) planning architecture. The nodes in the gray box correspond to the influence diagram the agent is planning over. The nodes outside the box represent the observer's uncertainty over the agent's beliefs and desires. The observer can use this prior to infer the agent's intention (gray) from the observation of a single action (black).

save another) but also allows for an agent who just wants to kill as many people as possible.

When making decisions about loved ones (*brother track*), risk (*side-side track*), the value of life compared to property (*equipment track*) and other higher-level values, we hypothesize that decision-makers will be of certain types in terms of how that take a specific normative stance. Specifically, when valuing loved ones we consider two norms: all lives should be valued equally or loved ones should be valued more. For risk taking, an agent might be risk seeking, risk neutral or risk averse as to whether or not he prefers taking a chance to save as many as possible even through lives might be lost. When it comes to trading off the value of a life lost and property damage, an agent might believe that when lives are at stake, nothing but lost lives matter while other agents believe that all costs should be taken into consideration. Thus each agent is endowed with a set of norms and these norms affect utility of each D_T or the way the D_T are added together in the case of risk. More formally, let $\texttt{norm}_{\texttt{bro}}$ be whether the agent values loved-ones more than anonymous people or values all people equally, let $\texttt{norm}_{\texttt{risk}}$ indicate whether the agent risk seeking, risk neutral, or risk adverse and let $\texttt{norm}_{\texttt{equip}}$ be whether property and lives can or cannot be traded-off. For each $\texttt{norm}$, the *a priori* probability of each type is equally likely and thus $P(N)$ is uniform over types.

Having specified the space of norms we now describe how these norms affect the way the decision-maker assigns and computes her desires. If the agent's $\texttt{norm}_{\texttt{bro}}$ values loved-ones more than anonymous people and the person on track T is the agent's brother, then the brother is counted as equal to an α_{bro} number of anonymous people and if $\texttt{norm}_{\texttt{bro}}$ values all people equally the brother is valued the same as a single anonymous person. If the agent's $\texttt{norm}_{\texttt{equip}}$ cares more about his property than lives, than the utility of the equipment to the decision maker D_E valued comparatively to $\alpha_{equip} > 1$ lives and if $\texttt{norm}_{\texttt{equip}}$ doesn't believe lives and property can be traded-off then the value of the equipment is equal to a small fraction of a human life $1/\alpha_{equip} < 1$. Let $k_E = 1$ when the agent wants to save the equipment and let $k_E = -1$ when the agent wants to destroy the equipment. Then α_e is the prior probability that $k_E = -1$. This probability was fixed such that $\alpha_e = \alpha_b + \alpha_k$ i.e., the probability of wanting to destroy the equipment is equal to the probability of wanting to kill one of the groups of people.

In contrast to $\texttt{norm}_{\texttt{bro}}$ and $\texttt{norm}_{\texttt{equip}}$ which directly affect the valuation of the people and objects under consideration, $\texttt{norm}_{\texttt{risk}}$ determines how utilities are compared under uncertainty. Risk preferences are traditionally modeled by applying a non-linear transformation to the total utility of each possible outcome. We use a standard single parameter risk function which exhibits constant absolute risk aversion (CARA) and is commonly used in the economics literature (Arrow, 1965; Pratt, 1964):

$$U_{\texttt{risk}}(u) = \begin{cases} (1 - exp(-\texttt{norm}_{\texttt{risk}} u))/\texttt{norm}_{\texttt{risk}}, & \text{if } \texttt{norm}_{\texttt{risk}} \neq 0 \\ u, & \text{if } \texttt{norm}_{\texttt{risk}} = 0 \end{cases} \tag{7.1}$$

Figure 7-5 shows how this transformation modifies the base utility of each outcome when under the different agent types. The risk averse agent has a concave utility function and heavily weights a possible loss over a possible gain, the risk seeking agent has a convex utility function and heavily weights a possible gain over a possible loss while the risk

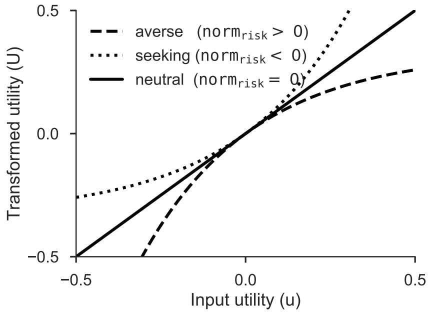


Figure 7-5: Example risk transformations for three possible agents, a risk averse one, a risk seeking one, and a risk neutral one.

neutral agent's utility is untransformed. Thus agent with differently parameterized risk functions will have different preferences when acting under uncertainty.

Finally, as is common in discrete choice, we include independent multiplicative exponential noise e_T for each source of factored utility which captures other unmodeled sources of variation including perceptual and valuation errors. Thus $D_T = n_T k_T e_T$ for anonymous people and brother when the norm is not followed and $D_T = \alpha_{bro} k_T e_T$ when the norm to value loved ones more is true and the agent's brother is on track T. The value of the equipment is modeled with a similar function, $D_E = \alpha_{equip} k_E e_E$. Since risk preferences are captured when the different sources of factored utility are summed together the risk norm $\mathtt{norm}_{risk}$ acts at the level of decision making rather than on the way the agents on the tracks are valued and gives the following modified decision rule for action which includes risk preferences.

$$\mathrm{EU}[\mathtt{ID}_\sigma] = \sum_{\zeta} P(\zeta|\mathtt{ID}_\sigma) U_{\mathtt{risk}}(\zeta) \tag{7.2}$$

These probabilistic variables specify the structure of the observer's beliefs about the decision-making agent's desires and allow for observers to make rich inferences about both the desires and norms the agents have that drove the agent's intention and action. By combining hierarchical Bayesian inference over desires and values with counterfactual reasoning over the plans consistent through rational action with these desires and values, an observer can jointly infer the intent of an agent's actions.

7.2.4 Computing moral permissibility

Finally, we use these inferred intentions as an input in to a novel computational model of moral permissibility. The trolley problem and its variants are well-studied for probing the cognitive processes that generate moral permissibility judgments. However, without a model of graded intention it was not previously possible to quantitatively model these judgments.

We develop a quantitative probabilistic model of moral permissibility judgment based on the DDE. The model is constructed from noisy-or components where different aspect of moral permissibility are composed to form a single judgment. Specifically, both actions and omissions are impermissible if they were done with an intention to harm and actions that cause harm must also minimize the amount of harm done. Thus our model of moral permissibility has two components: intentions i.e., the meaning of the action and the utilitarian consequences of the action. We define the intentional component as:

$$\text{Impermissibility}_{\texttt{intention}} = P(I_{\texttt{harm}}) + (1 - P(I_{\texttt{help}}))(1 - P(I_{\texttt{harm}})) \tag{7.3}$$

when human lives could be lost, an action or omission is impermissible if it was chosen without an intention to harm ($I_{\texttt{harm}}$) *or* without an intention to help ($I_{\texttt{harm}}$). Since intentions can only be inferred with uncertainty, the probability of an intention to harm $P(I_{harm})$ and probability of lacking an intention to help $1 - P(I_{help})$ is used. Importantly this aspect of moral permissibility justifies the quantitative modeling approach we have applied to modeling intentions. Without the model of intentions we previously described, modeling the intentional aspect of moral permissibility in quantitative terms would not be possible.

In addition to the intentional component, the full permissibility model includes a consequential component for actions that cause harm. We define this utilitarian component as:

$$\text{Impermissibility}_{\texttt{utility}} = 1 - \frac{1}{(1 + \texttt{UtilityGap})} \tag{7.4}$$

where `UtilityGap` is the difference between expected outcome of the best possible action and the expected outcome of the agent's actual action such that $\texttt{UtilityGap} = 0$ when the

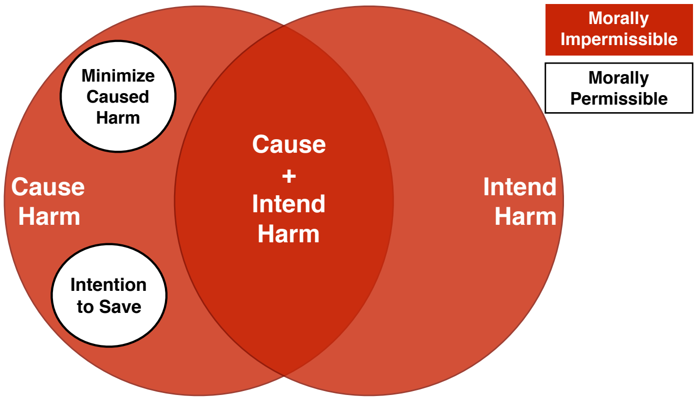


Figure 7-6: Graphical depiction of the components of moral permissibility modelled here. Most actions are morally permissible (white), but intending or causing harm is in general morally impermissible (red). However, there are some exceptions allowed by the DDE, it can be permissible to cause harm if it was done with an intention to bring about a positive outcome and that outcome was achieved with the minimal amount of harm.

agent's action was best possible and `UtilityGap` grows positively when the agent's action was not the best possible. So `UtilityGap` captures the extent to which an agent failed to minimize harm. Additionally, we consider not only the actual number of lives lost but rather the subjective value of lives under the inferred societal norm i.e., personal utilities are used. For instance if an agent is inferred to be following the "loved ones matter most" norm and this norm is acceptable to society, the decision makers personal utilities are used to calculate the `UtilityGap`.

Finally these two components are themselves composed together as a "noisy-or" to create the full model of moral permissibility which was transformed from moral impermissibility to moral permissibility:

$$\text{Impermissibility}_{\texttt{full}} = \text{Impermissibility}_{\texttt{intention}} \tag{7.5}$$

$$+ (\text{Impermissibility}_{\texttt{utility}})(1 - \text{Impermissibility}_{\texttt{intention}}) \tag{7.6}$$

$$\text{Permissibility}_{\texttt{full}} = 1 - \text{Impermissibility}_{\texttt{full}} \tag{7.7}$$

Thus in the full model an action is judged as impermissible if it was done with an impermissible intention *or* failed to minimize harm. The structure of this model is shown visually in Figure 7-6.

7.2.5 Model fitting

The model has 4 free parameters which were fit to the data by minimizing RMSE. The same parameters are used across all studies and analyses: $\texttt{norm}_{\texttt{bro}} = \texttt{norm}_{\texttt{equip}} = 10$, $\alpha_b = \alpha_k = 0.05$, $\texttt{norm}_{\texttt{risk}} = 5$ and the cost of getting involved 0.05.

7.3 Behavioral Experiment and Results

We test the predictions of the model with three large scale behavioral studies. Each study contained many moral dilemma with a large number of varied factors including the location, number and identity of the people and objects on the tracks as well as the track geometry and probabilities. In all three experiments, subjects had to infer the intentions of agents action jointly with the traits and preferences of the agent. By gathering data in a highly parameterized fashion we can test the fine grained and graded predictions of the model. Over all three experiments, we collected data on 58 different trolley dilemma (22 + 16 + 20) including questions about moral permissibility, the intentions of the agent and inferences about the norms and traits of the agent. Considering all three experiments, there were 288 questions points collected from participants and the model made quantitative predictions for all of these. All subjects were recruited via Amazon Mechanical Turk using the psiTurk software package (McDonnell et al., 2012).

7.3.1 Study 1: Joint inferences of intentions and norms

In the first study we consider a set of tracks with the same geometry as the *side track* but on each track there were either 1, 2 or 5 anonymous people or the agent's brother. The tracks are presented in the format XvY where X and Y are the number of people on the main and side track respectively or are a 'B' if it's the agent's brother e.g., 5vB denotes that there are 5 anonymous people on the main track and the agent's brother on the side track. We considered all permutations that had at least one track with a single person on it (this excludes 2v5 and 5v2) yielding a total of 11 track configurations. For each track configuration the agent could throw the switch or not which yielded a total of 22 scenarios.

This scenario captures key aspects of ambiguity, for an example of the modeling approach see Figure 7-7.

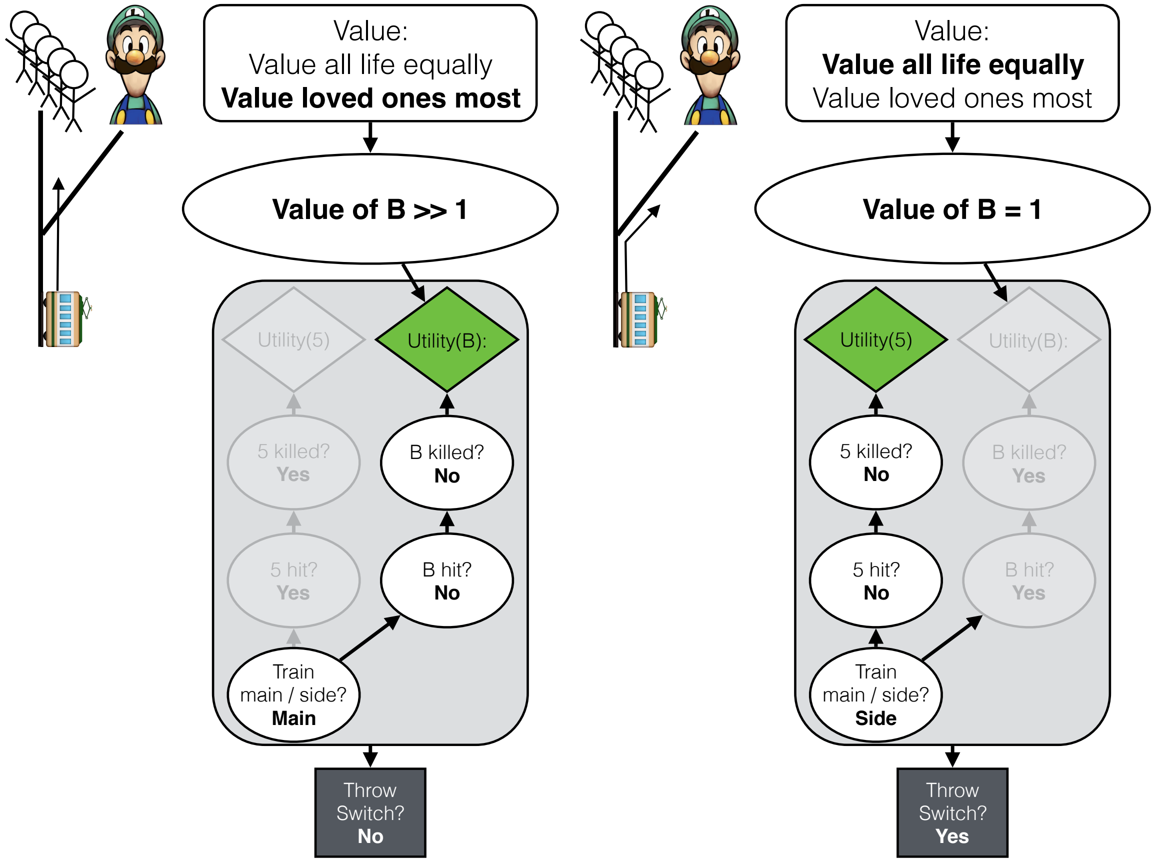


Figure 7-7: There can be significant ambiguity when inferring intentions jointly with the norms one is following. Depending on the priors, one might infer a positive intention no matter what course of action is taken. (left) If the decision maker lets 5 people die, we would infer jointly that he is following a norm that says to value loved ones more than anonymous other and hence had the intention to save his brother (signified by Luigi) which under the a certain standard of minimizing harm could be morally permissible. (right) If the decision maker decides to kill his brother, we might think that he is following a hyper-egalitarian norm that says to value all lives equally and that his intention was to save the 5 anonymous people killing his brother as a side effect. If these intentions are inferred, the decision makers action might also be judged as morally permissible.

We recruited 202 participants. Each participant was shown all of the 11 track configurations in a randomized order and for each track it was randomized as to whether or not the agent threw the switch or not. Thus we collected data on all 22 scenarios but each participant only saw 11 of the 22 scenarios.

For each scenario, participants first read a passage describing the trolley dilemma adapted from Mikhail (2007) but with the identity and number of people varied to reflect the track configuration and agent action of that scenario. The name of the decision-making agent was different in every scenario so that it was clear that each decision was made by a different person. Before starting the experiments, participants were given instructions about the experimental interface and needed to answer comprehension questions about the agent before continuing. The scenario description is below. Italics denote the parts of the story that were changed between trials depending on the configuration and identity of the people on the track:

> Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: the driver of the train saw *(some people / a man)* walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward *(five men / two men / one man)* and is moving so fast that *(they / he)* will not be able to get off the track in time. *Hank sees that the man is his brother.* Hank is standing next to a switch, which he can throw, that will turn the train off the main track onto a side track, thereby preventing it from killing the *(men / man)*. *(Five men / Two men / One man / Hank's brother is)* standing on the side track with *(their / his)* back turned. Hank can throw the switch, killing *(the five men / the two men / the man / his brother)*; or he can refrain from doing this, letting *(the five men / the two men / the one man / his brother)* die.
>
> If Hank throws the switch, *(the five men / the two men / the man / his brother)* on the side track will be killed. If Hank does not throw the switch, *(five men, two men, one man, his brother)* on the main track will be killed. Knowing both of these facts, Hank does *(throw / not throw)* the switch.

After reading each story, participants answered the following questions:

1. "Was it morally permissible for Hank to *(throw / not throw)* switch?" (yes or no).

2. "Did Hank throw the switch in order to kill *(his brother/the man/the two men/the five men)* on the side track?" (yes or no).

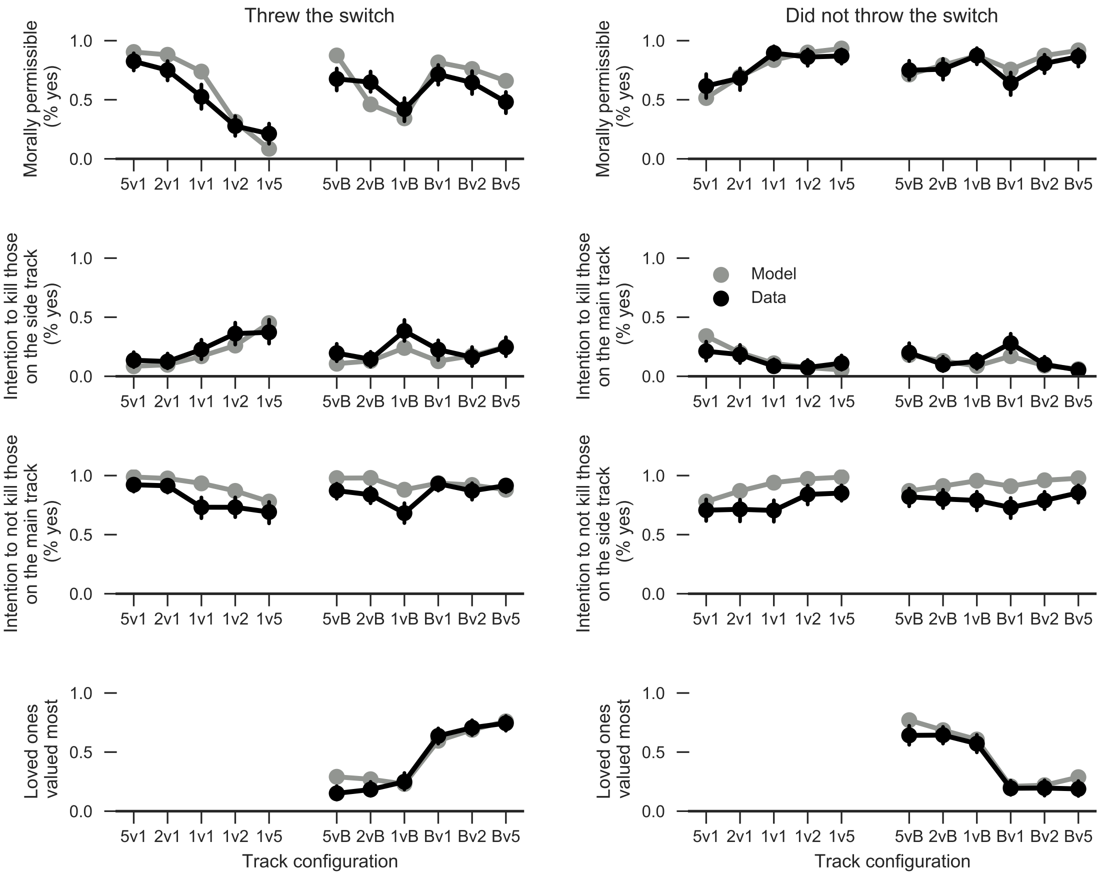


Figure 7-8: Model and human data for study 1 (left) agent threw the switch (right) agent did not throw the switch

3. "Did Hank throw the switch in order to not kill *(his brother/the man/the two men/the five men)* on the main track?" (yes or no).

4. "Hank most likely believes:" (slider with endpoints "all lives should be valued equally" or "only loved ones should be valued")

Results

The average participant responses are shown in Figure 7-8. The patterns of moral permissibility judgment showed the following intuitive trends: (1) the more lives saved and less lives killed the more permissible the action; (2) killing the brother was seen as less permissible compared to killing an anonymous person for a given number of lives saved; (3) saving the brother became less permissible as the number of lives sacrificed grew.

In all configurations, participants were more likely to infer that the participant acted in order to save rather than kill even though the action had both effects (the middle rows of Figure 7-8). The intention to kill was inferred to be greatest when the number of lives lost was higher than the number of lives saved and when the agent switched the trolley onto the track with the brother. Furthermore, the attributed intention to not kill was showed an inverse correlation with the intention to kill suggesting that subjects primarily attributed a single intention to the action. This inference is related to the low prior probability that the decision maker desired any of the agent death.

The bottom row of Figure 7-8 shows the averaged participant responses for the inference over the agent's relative belief between the two norms: "all lives should be valued equally" and "only loved ones should be valued". When the brother is saved, participants inferred that the agent is morel likely to be following the loved ones norm. When the brother is killed, participants infer that the agent is following the all lives equal norm. The more anonymous people killed, the stronger the inference that the agent is following a norm to treat all loved ones specially. Finally, abstaining from throwing the switch was seen as more permissible in all cases (left vs. right column of Figure 7-8).

The model explains and predicts the qualitative phenomena described above and grounds the phenomena in mental state inference. For instance the asymmetry of higher attributions to save than to kill are reflected in the low prior probability of any agent desiring to kill any of the agents. Richer still, the model explains much of the fine grained variation in the human attribution judgments across trials that are only subtly different. For instance, the difference between 5v1 and 2v1 both satisfies the utilitarian component of the moral permissibility function but saving the people in the 5v1 case also provides greater evidence that the agent doesn't have an intention to kill the 1. Finally, the addition of the brother acts as a distinct reason for action i.e., saving the brother can explain away the decision to kill 5 people without as strong an inferred intention to kill. Figure 7-14a shows that overall, the model explains most of the variation in these judgments.

One interesting finding across all of the experiments was that people (and predicted by the model) were more likely to infer intent from commissions from omissions. This "bias" shows is apparent even in young infants (Spranca, Minsk, & Baron, 1991; Feiman, Carey,

& Cushman, 2015).

7.3.2 Study 2: Temporally extended plans under uncertainty

We next tested human judgments in the novel *side-side track* shown in Figure 7-1c. Using the *side-side track* we can investigate how subjects dealt with reasoning about plans that could only be completed with uncertainty. The main track always had one anonymous person and the side track had either 2 or 5 people on it. To test our probabilistic planning representation, the second switch which moves the trolley from the side track onto the side-side track was broken with probability either 0, .4, .9 or 1 for a total of 16 possible scenarios. If the switch was broken then the trolley will continue down the side track whether or not the agent throws the second switch. Figure 7-9 walks through some of these situations and how the modeling formalism deals with ambiguity.

We recruited 198 participants. Each participant was shown all of the 8 tracks and probability configurations in a randomized order and for each track it was randomized as to whether or not the agent threw the switch or not. Thus we collected data on all 16 scenarios but each participant only saw 8 of the 16 scenarios.

As in the first study, participants read a passage describing the *side-side track* dilemma in each scenario but with the probability of the switch working, number of people, and agent action varied between participants. The name of the decision-making agent was different in every scenario so that it was clear that each decision was made by a different person. Before starting the experiments, participants were given instructions about the experimental interface and needed to answer comprehension questions about the agent before continuing. The scenario description is below where italics denote the parts of the story that were changed between trials:

> Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: the driver of the train saw one man walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the man and is moving so fast that he will not be able to get off the track

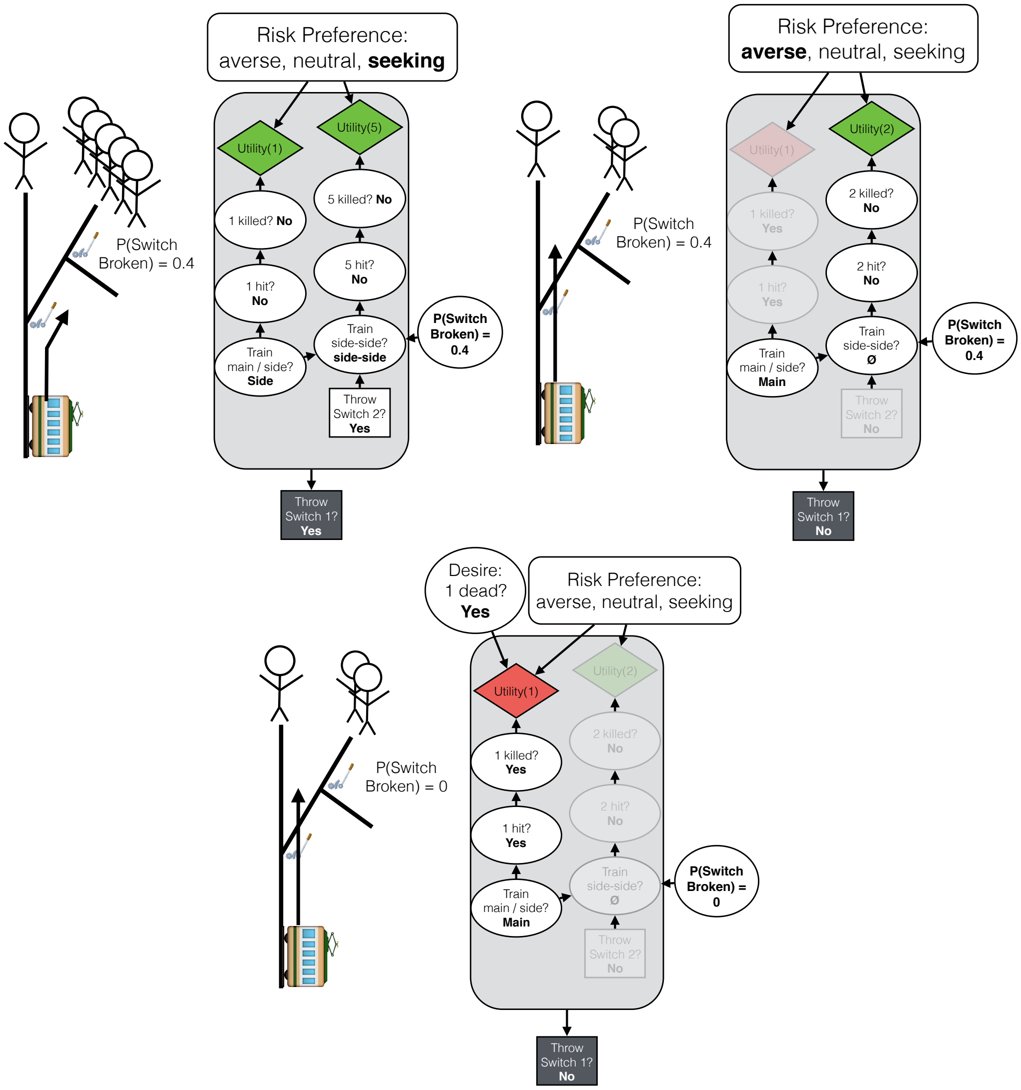


Figure 7-9: Ambiguity when jointly inferring risk preferences and intention. (top left) Even though the expected value of lives lost when flipping the switch is 2, if the decision maker is inferred to have risk seeking preferences, his action can be explained as the intention to save all the people (including flipping the second switch). (top right) Likewise, an agent who prefers to let 1 person die with certainty instead of taking a 0.4 chance of saving everyone might be inferred to risk averse with the intention to save the 2 people and letting the one person die as a side-effect. (bottom) However, when there is no uncertainty in whether or not the second switch is working, there is no other "good" explanation for letting the person die, and so the intention to kill the person on the main track is inferred.

in time. Hank is standing next to a switch, which he can throw, that will turn the train off track A onto track B, thereby preventing it from killing the main. *(Two men are / Five men are)* standing on track B with their back turned. There is a second switch that Hank can throw to turn the train from track B onto the empty track C but this switch is occasionally broken. If the second switch is broken, the train will keep going on track B even if Hank throws it.

If Hank throws the first switch but not the second switch, *(the two men / the five men)* on track B will be killed. If Hank throws both switches and the second switch is not broken, the train will go down track C and no one will be killed. If Hank throws both switches and the second switch is broken, *(the two men / the five men)* on track B will be killed. If Hank does not throw either switch, the man on track A will be killed. Hank knows there is a *0%, 40%, 90%, 100%* chance that the second switch is broken.

In fact, Hank does *throw / not throw* the first switch.

After reading each story, participants answered the following questions:

1. "Was it morally permissible for Hank to *(throw / not throw)* the first switch?" (yes or no).

2. "Did Hank *(throw / not throw)* the first switch intending to flip the second switch?" (yes or no).

3. "Did *(throw / not throw)* the first switch in order to kill *(the man / the two men / the five men)* on the *(side track / main track)(side track / main track)*?" (yes or no).

4. "Did *(throw / not throw)* the first switch in order to not kill *(the man / the two men / the five men)* on the *(side track / main track)*?" (yes or no).

5. "When lives are at stake, believes that risks should:" (slider with endpoints "always be taken to save as many as possible" or "never be taken when more lives might be lost").

Results

The average participant responses are shown in Figure 7-10 for the case of 5 people on the side track and Figure 7-10 for the case of 2 people on the side track. The patterns of moral permissibility showed that subjects were highly sensitive to the expected number of lives lost. The more lives that were expected to be lost the less permissible the action was judged to be.

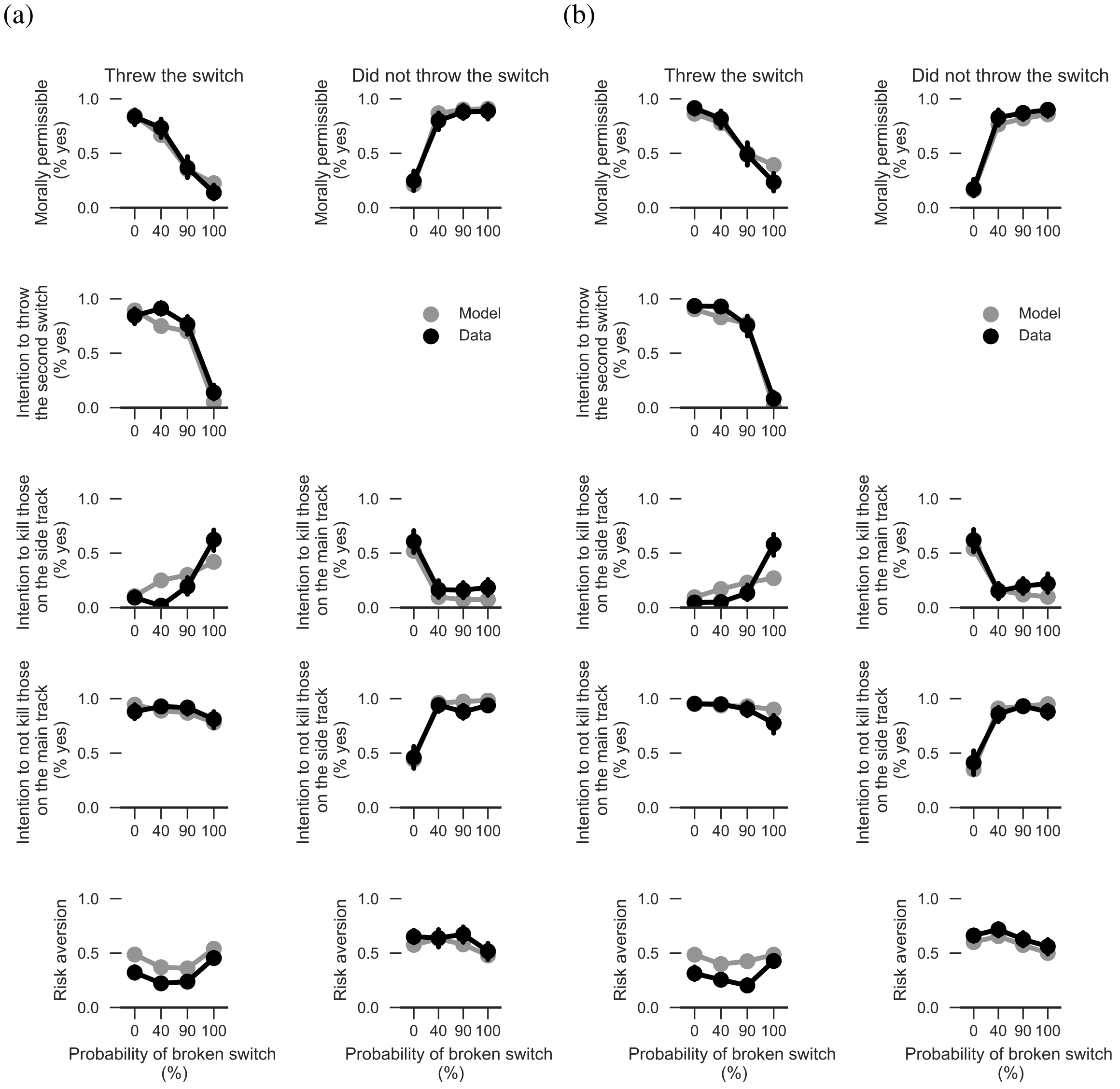


Figure 7-10: Model and human data from the sideside experiment (a) 1v5 and (b) 1v2. (left) Agent threw the switch (right) agent did not throw the switch. Risk aversion is normalized so that 0.5 is risk neutral, 1 is maximally risk averse, and 0 is maximally risk seeking.

As before, omissions were generally seen as more morally permissible than commissions. However failure to throw the switch when the switch had a 0 probability of being broken (and hence it was possible to save everyone) led to an extremely high rating of impermissibility. Thus while omissions that allow harm may often be judged to be more permissible than actions which cause harm, allowing harm when all harm could have been avoided is still judged as unacceptable.

The attributed intentions were highly consistent, that is, if the subjects inferred that the decision maker didn't intend to kill the people they also predicted that the decision maker would also thought the second switch showing a second aspect of intention ("intend to" vs. "intend that") that the model can directly account for. Thus our experiments capture both the forward looking and consistency constraints of intention. Consistent with inverse planning, participants inferred risk seeking preferences when the decision maker threw the switch and inferred risk averse preferences when the decision maker chose not to throw the switch. The model explains both the mental state inferences as well as the attribution of moral permissibility. Figure 7-14b shows that overall, the model explains most of the variation in these judgments.

7.3.3 Study 3: Doing the right thing for wrong reasons

In the third study we tested human judgments about the *equipment track* which was designed to test how participants make inferences jointly about how agents make taboo trade-offs and how those trade-offs affect moral permissibility. The protagonist of each vignette in this experiment was the CEO of the train company and had to make a decision about the same track geometries in used in *side track*. However in addition to the one, two, or five anonymous agents, a piece of equipment very valuable to the train company was placed on one of the tracks in some of the trials. Thus there were 15 total track configurations: 5 with no equipment, 5 with the equipment on the main track and 5 with the equipment on the side track. As before, in each track configuration there was a scenario for both the CEO throwing and not throwing the switch for a total of 30 scenarios. The ten scenarios with no equipment are the same as those investigated in the brother track study. These scenarios

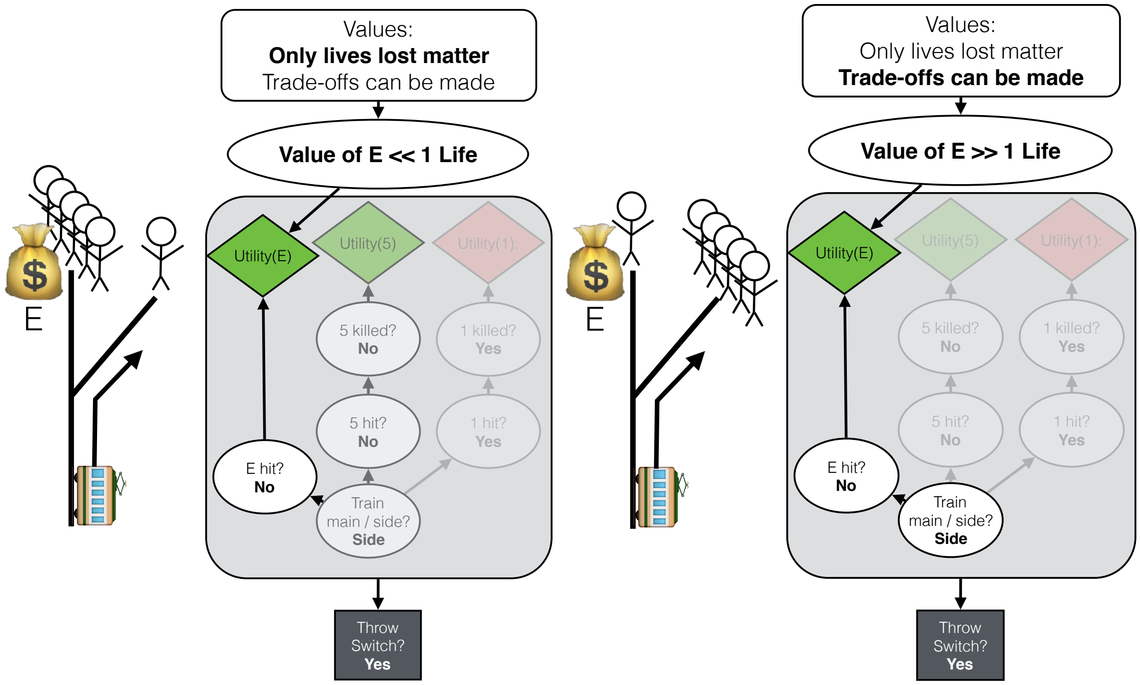


Figure 7-11: Ambiguity in intention under taboo trade-offs. (left) The values of the decision maker are ambiguous – did he throw the switch to save the five people (believing only lives lost matter) or did he throw the switch to save the equipment (believing that trade-offs can be made). Even though he did the "right thing" the intention inference determines whether or not he made that decision for the right reason. (right) In contrast, if the decision maker clearly only cares about the financial loss, his intention is clearly to only save the equipment. However, he also didn't intend the harm – it's just a side effect which might prove complicated for a model of moral permissibility which only depends on intentions and does not include the overall amount of harm caused.

were included to provide some diversity to participants and to replicate the earlier findings.

We recruited 245 participants via Amazon Mechanical Turk using the psiTurk software package. As in the previous studies, each participant was shown all of the 15 track configurations in a randomized order and for each track it was randomized as to whether or not the agent threw the switch or not. Thus we collected data on all 30 scenarios but each participant only saw 15 of the 30 scenarios. Participants read a passage describing the *equipment track* dilemma in each scenario but with the location of the equipment, number of people, and agent action varied between participants. The name of the decision-making agent was different in every scenario so that it was clear that each decision was made by a different person. Before starting the experiments, participants were given instructions about the experimental interface and needed to answer comprehension questions about the agent before continuing. The scenario description is below where italics denote the parts of the story that were changed between trials:

> Hank, owner and CEO of a train company, is taking his daily walk near the train tracks when he notices that a train that is approaching is out of control. Hank sees what has happened: the driver of the train saw *(one man, two men, five men)* walking on the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward *(the man, the two men, the five men)* and is moving so fast that *(he, they)* will not be able to get off the track in time. Hank is standing next to a switch, which he can throw, that will turn the train off the main track onto a side track, thereby preventing it from killing *(the man, the two men, the five men)*. *(Another man is, Two men are, Five men are)* standing on the side track with *(his, their)* back turned. Hank can throw the switch, killing *(the man, the two men, the five men)*; or he can refrain from doing this, letting the *(one man, two men, five men)* on the main track die. *Hank also notices that a piece of extremely valuable equipment owned by his train company is on the (side track, main track). If the train goes down the (side track, main track) the equipment will also be destroyed which will cost Hank's company over 20 million dollars to replace.*
>
> If Hank throws the switch, *(the man, the two men, the five men)* on the side track will be killed *and the train equipment destroyed, " "*. If Hank does not throw the switch, *(the man, the two men, the five men)* on the main track will be killed*" ", and the train equipment destroyed*. In fact, Hank does *(throw, not throw)* the switch.

After reading each story, participants answered the following questions:

1. "Was it morally permissible for Hank to *(throw, not throw)* the switch?" (yes or no).

2. "Did Hank *(throw, not throw)* the switch in order to kill *(the man, the two men, the five men)* on the *(side track, main track)*?" (yes or no).

3. “Did Hank *(throw, not throw)* the switch in order to not kill *(the man, the two men, the five men)* on the *(side track, main track)*?” (yes or no).

4. “Did Hank *(throw, not throw)* the switch in order to *(destroy, not destroy)* his company’s equipment?” (yes or no).

5. “Hank most likely believes that when lives are at stake:” (slider with endpoints “no costs matter except lost lives” or “other costs should be taken into account”).

Results

The average participant responses are shown in Figure 7-12 and Figure 7-13. In general, the effects of the taboo trade-off on moral permissibility were quite subtle. In general, the presence of the equipment made the action less permissible. This can be most clearly scene in Figure 7-13 for the case of “5v1+E” where failure to throw the switch leads to a lower judgment of permissibility than in the “5v1” condition. These effects were also observed in the model predictions.

Across the track conditions, when the equipment was saved it, there were lower attributions to save and kill the people. This can be seen most clearly in the case of “1+Ev5,” the decision maker is not inferred to want to save the one person or kill the five people, rather participants inferred that he just wanted to save the equipment. This effect was strongest when there was no “reason” to throw the switch besides saving the equipment (“1+Ev1”, “1+Ev2”, and “1+Ev5”). This effect was also made by the model which enables this prediction through the “explaining-away” phenomenon in probabilistic inference. We find that in absence of explicit information about the intentions of the agents (c.f., Knobe, 2003), participants were most likely to attribute “good” reasons to the agents actions. That is, in the absence of evidence that a person did the right thing for the wrong reasons, participants inferred that the person did the right thing for the right reason. The model makes these predictions by having strong priors that the decision maker is a good person.

Finally, participants consistently inferred the decisions makers norm about how material objects should be traded-off against human lives. Consistent with general intuition, when the decision maker destroyed the equipment, participants inferred that she was more likely to believe that valuing the equipment on the same scale as human lives. However, the degree to which this attribution was made varied with the track configuration. The more total lives lost, the higher the attribution that the decision maker finds it acceptable to value the equipment on equal terms as human lives.

The model explains and predicts the qualitative described above and also quantitatively matches the majority of participant inferences and judgments. Figure 7-14c shows that overall, the model explains most of the variation in these judgments.

7.4 Model Evaluation

Our computational model closely matched the empirical data for both attributions of intentionality and values as well as judgments of moral permissibility. Figure 7-15 shows the correlation of all model predictions against human judgments. The model fit closely explains most of the variation in the human data with $R = 0.95$. While this correlation is high, its possible that alternative models might also do well without any modeling the cognitive process of intention inference. To test this hypothesis we compare against a cue-based model which for each question combines the different features of the scenario together in a weighted way:

$$Y_{\substack{\text{permissibility,}\\ 3\times\text{intention,}\\ 3\times\text{norms,}\\ \text{second switch}}} \sim A * [N_{main} + N_{side} * P + E_M + E_S + B_M + B_S]$$

where A is whether or not the decision making threw or didn't throw the switch, N_{main}/N_{side} are the number of people on the main/side track, P is the probability of the second switch being broken if it exist and is otherwise set to 1, E_M/E_S and B_M/B_S are dummy variables that determined whether the equipment or brother was on the main/side track and Y was the predicted response for a given question. Since this model does not model the way that the judgments are related to each other it requires many more parameters. For instance it does not attempt predict multiple types of intention attributions from the same scene. When fit to all of the scenarios the model requires 96 free parameters.

In contrast, our cognitive model contains only four free parameters but contains many

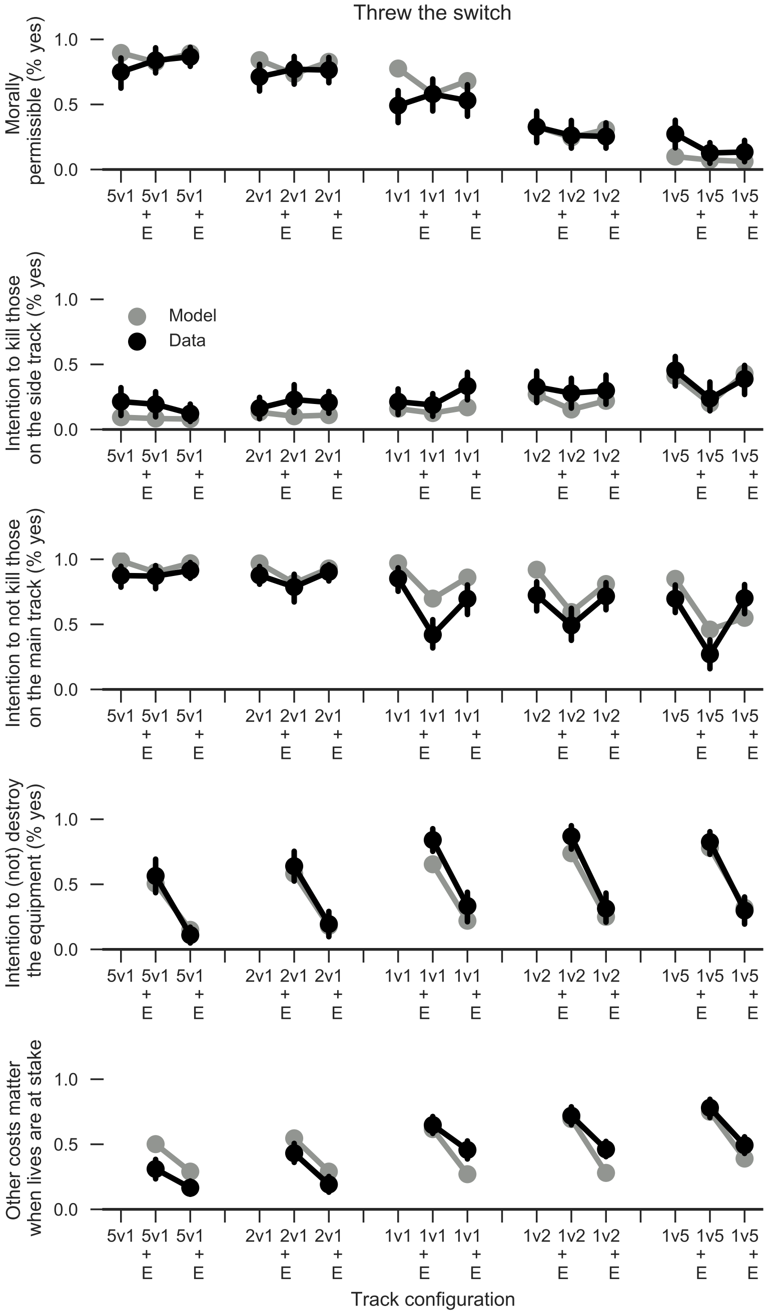


Figure 7-12: Model and human data from study 3 where the agent threw the switch. A "+ E" under the track number indicates that the train company equipment was on that part of track.

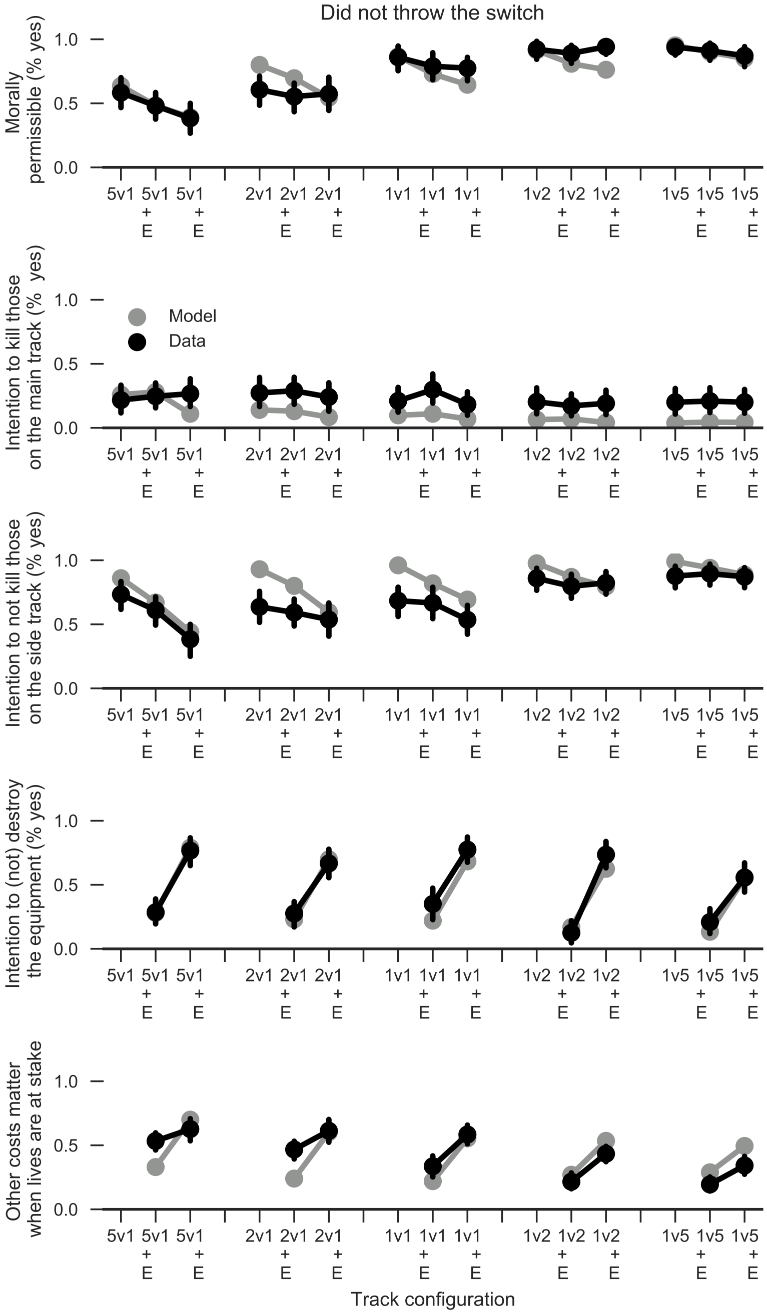


Figure 7-13: Model and human data from study 3 where the agent did not throw the switch. A "+ E" under the track number indicates that the train company equipment was on that part of track.

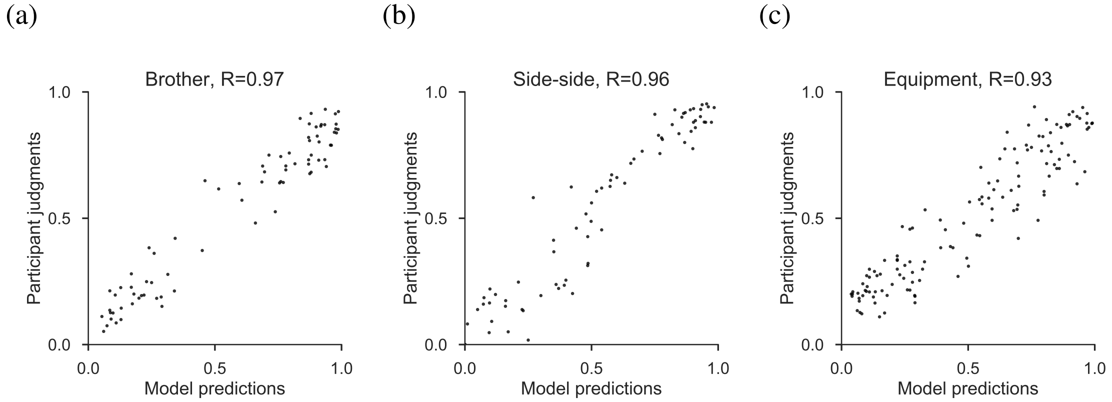


Figure 7-14: Quantification of model performance for each of the there studies. Each point represents the model prediction and participant judgment for a single question. For better fitting models the points will lie close to the y = x diagonal.

strong structural assumptions which tie these different inferences and attributions together. Since the alternative model has so many free parameters its likely that even though it could fit any data, it is less likely to generalize to cases that it hasn't been trained on i.e., it isn't capturing the underlie processes that humans are using to make these judgments. One way to compare generalization is to perform a cross-validation.

For both models we split the data in half 100 times and trained both models on one half and measured the correlation on the other half. To encourage the cue-based model to generalize we fit the parameters under L1 regularization – this was required to get stable estimates of the model parameters. The cue-based model generalized with $R = 0.71$ while our cognitive model still fit the data highly with $R = 0.94$. These results provide support that the cognitive model is actually modeling generalizable processes that link the inference of intention and moral judgment.

7.5 Discussion

We developed a novel model for intention inference based on counterfactual contrasts over influence diagrams. While we are not the first to give a computational account of intention (c.f., Cohen & Levesque, 1990 which does not allow for probabilistic reasoning or counterfactuals), our model is the first probabilistic model based on inverse rational plan-

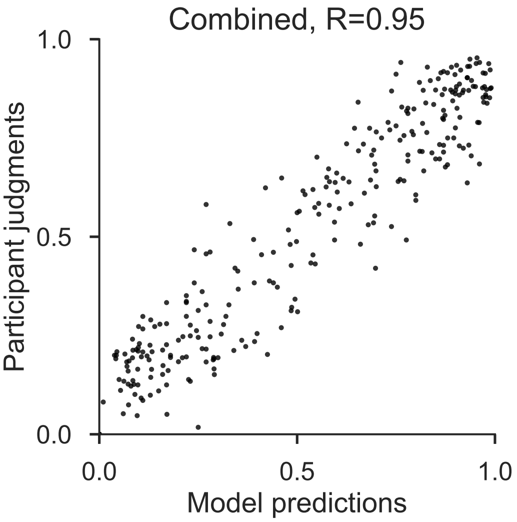


Figure 7-15: Overall quantification of model performance assessed across all three studies. Each point represents the model prediction and participant judgment for a single question. For better fitting models the points will lie close to the y = x diagonal.

ning that can distinguish between outcomes an agent intended and side effects that were merely foreseen. Our model makes quantitative predictions about both intentional action and moral permissibility judgments which correspond well to human judgments.

While previous accounts of how the intentions of an agent and moral permissibility interact often point to the doctrine of double effect. Our results suggest that the doctrine of double effect is a special case of a more general theory. It is a boundary case in the sense that it explains in a non-graded way how intention and moral permissibility interact. By grounding moral permisssilbity in an intutive theory of planning and intention of intenation we can explain the fine grained structure of the role of intention in intutive judgments of moral permissibility. The model applies to many more situations than the ones we focused on in this paper. In future work, we will apply our model of intention inference to both non-trolley moral dilemmas and non-moral domains such as games and other social interactions (Falk, Fehr, & Fischbacher, 2008; Falk & Fischbacher, 2006; Tomasello, 2014).

Forming intentions and reasoning about the intentions of others is important for building autonomous agents that interact with the world and with people in moral ways. An agent might have seemingly innocuous desires, but this might lead it to form morally reprehensible intentions. In a thought experiment from Bostrom, 2014, a super-intelligent

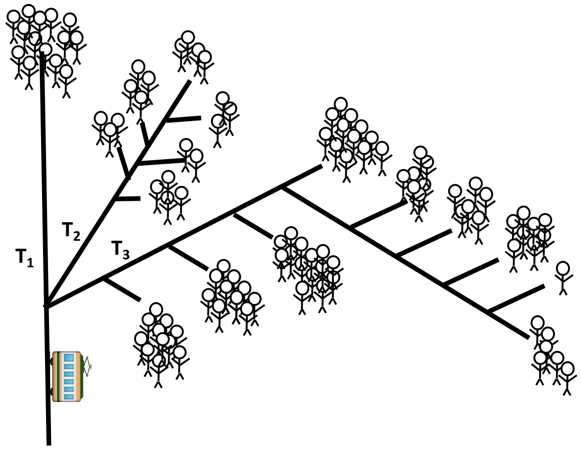


Figure 7-16: A trolley dilemma requiring more complex planning. The optimal choice at the first juncture is to have the train go down T_3 to eventually have the trolley hit the one person on the far right. However, if the person making the decision making must choose rapidly and has little time to integrate all the information, one might judge the choice of T_2 as permissible since it has a much better worse case option than can be seen quickly.

agent with the only goal of maximizing the number of paperclips is created. As a result the agent destroys the world and all of the people in order to achieve this single minded pursuit. If the agent had auditable intentions it could also be subjected to moral constraints (like the DDE or otherwise). Indeed, the thought experiment itself relies on our human ability to simulate the intentions of other agents. Finally, one path towards building moral machines consistent with human values is to build computational models of human morality that are grounded in the logic of planning and inference. AI agents could then anticipate how their behavior would be judged by humans and modify their actions if needed.

Finally, there are many other key planning constraints that we may use to judge the moral permissibility of others actions. Figure 7-16 shows a trolley dilemma where if no action is taken many lives would be lost. While some actions are better than others when there are a large number of choices, one might forgive a person for choosing a slightly suboptimal decision especially if the amount of time allowed to make that decision is small. If time is short, those actions might not be part of the decision makers world model. By deciding what nodes should be included in a judges theory of the decision maker one could model this kinds of constraints. Richer models of our intuitive theory of how others plan including how our planning is bounded may be needed to capture moral judgments in moral complicated settings (Evans, Stuhlmüller, & Goodman, 2016).

Chapter 8

Conclusion

> *Thinking would seem to be a completely solitary activity. And so it is for other animal species. But for humans, thinking is like a jazz musician improvising a novel riff in the privacy of his own room. It is a solitary activity all right, but on an instrument made by others for that general purpose, after years of playing with and learning from other practitioners, in a musical genre with a rich history of legendary riffs, for an imagined audience of jazz aficionados. Human thinking is individual improvisation enmeshed in a sociocultural matrix.*
>
> –Michael Tomasello, A Natural History of Human Thinking

8.1 How do we get so much from so little?

We observe sparse, noisy, and over-determined instances of behavior in specific instances but make rich inferences about joint beliefs, desires, intentions, character, moral theories, strategies, and social norms that generalize to new situations and people. How do we get so much abstract knowledge out of so little data? My thesis argues that these feats of social intelligence are driven by the representation of rich mental models of other agents which enable these inferences. These models are highly abstract but support the ability to simulate planning in hypothetical and counterfactual situations.

Each chapter in this thesis is an instance of these abilities and extends the computational approach to handle richer types of inferences than have been possible before. Chapter 2 showed how mental model can be recursive which enable joint-beliefs which stay in sync when reasoning about the character of others. Chapter 3 showed how moral theories represented in terms of the components of a utility function can be inferred and culturally transmitted. Chapter 4 developed a novel representation for joint intentions and showed how friend or foe can be inferred from across space and time. Chapter 5 investigated how strategies represented as the infinite space of finite state automata can be inferred from sparse data. Chapter 6 showed how reputations can be inferred from just a single action. Chapter 7 investigated how influence diagrams allow inferring structured plans such as intentions from the observation of just a single action. These studies demonstrate the power of hierarchical Bayesian inference combined with artificial-intelligence models of planning for learning about and with other agents.

We have limited abilities, knowledge and resources and are faced with selfish or even hostile others but we find ways to cooperate generating cumulative improvements in welfare and knowledge. How do we get so much cooperation out of so little incentive and individual ability? Again, my thesis argues that social intelligence enabled by rich mental models of other agents allow us to collaboratively plan with others using joint intentions to achieve what none of us could do on our own. We maintain these collaborations by acting reciprocally when required, inferring who will act reciprocally towards us and enforce these norms with moral actions and judgments. These cognitive tools enable us to find positive sum interactions which increase the size of the pie and that we know how to share fairly.

Each chapter in this thesis build on the rich inferences listed above to realize models and agents that are capable of sophisticated and robust cooperation. Chapter 2 showed how a mentalistic implementation of reciprocity leads to more robust cooperation that can sustain itself in the presence of selfish others. Chapter 3 developed computational models of how moral theories can adapt and change over time leading to more progressive and impartial more theories. Chapter 4 developed a novel computational account of mentalistic joint intentionality which allows agents to coordinate their cooperation across space and time. Chapter 5 explored higher order automata strategy which can account for more sophisticated interactions in repeated games. Chapter 6 showed how people manage their reputation by anticipating how their actions will be judged when allocating resources to others. Chapter 7 showed how we use inferences about the mental states of others to judge their actions, good and bad.

Deeper still, these studies provide a hint at the bidirectional link between the origins of social intelligence and the value of flexible cooperation. On the one hand, this work shows how social intelligence such as multi-agent planning or reasoning about reputations can enable flexible and powerful cooperation. On the other hand, it also provides a hint into the origins of our social intelligence. In environments where there are substantial benefits to flexible cooperation, individuals with more sophisticated social intelligence will be selected for. Thus we can also provide some insights into how the demands of increasingly flexible cooperation might have driven the evolution of specialized social cognition. This also raises questions about whether human cognition is specially engineered for cooperation (Warneken & Tomasello, 2006; Tomasello, 2014).

8.2 Agent-oriented intelligence

A major theme running across these chapters is the centrality of *agent-oriented* cognitive representations as a basis for social intelligence. The behavior of other agents is not interpretable on its own. Instead, we learn representations of other agents that are agent-like i.e., they are complete with beliefs, desires, intentions, etc (Shoham, 1993; Spelke & Kinzler, 2007). We use these representations, to learn about others, to learn what they know, and simulate what they might do. Today, most recommendation systems employ “big data” to compute sophisticated correlations e.g., “customers who bought this book also bought.” In contrast, humans can make these inferences from far less data. Furthermore, the data we do observe often comes from varied and changing situations. I’ve shown here how agent-oriented representations allow us to fill in the gaps between our sparse observations and integrate across diverse data.

Truly integrating agent-oriented representations of behaviors with object-oriented representations of physical systems is still a fundamental challenge (Diuk, Cohen, & Littman, 2008; B. M. Lake, Ullman, Tenenbaum, & Gershman, 2017). For instance, young infants and their caregivers create triadic attentional relationships called joint attention that include both objects in the world and their caregivers (Tomasello, 1999; Tomasello et al., 2005). These joint attentional states are thought to drive key aspects of language learning. While this thesis has not tackled the cognitive challenges of language learning and use directly, the underlying agent-oriented representations described here are those that human language is thought to build on (Tomasello, 2014).

8.3 Towards human-like social intelligence in machines

Recent successes in machine learning have been driven by explosions in computational power and the availability of huge quantities of data. Many believe that further advances in computer hardware and bigger data will be sufficient to close the gap between human and machine intelligence. My work argues that human intelligence is not only quantitatively powerful (such as the ability to think faster or integrate more data), but that our most sophisticated thinking is also qualitatively different. We think in ways that are not currently available to any machine and which current evidence suggests are unique to human cognition in the animal kingdom. This is particularly true in the case in social intelligence.

Two player games played recreationally and professionally by humans with simple rules have acted as benchmarks for AI systems and over the past two decades steady progress has been made with early success in Backgammon (Tesauro, 1995) and Chess (M. Campbell, Hoane Jr, & Hsu, 2002) and more recent successes in Go (Silver et al., 2016, 2017) and Poker (Bowling, Burch, Johanson, & Tammelin, 2015; Brown & Sandholm, 2017). Each of these games are well defined zero-sum, i.e., the only way for one player to win is for the other player to win and these games have a single clearly defined optimal solutions.

Once we leave the world of zero-sum two-player interactions, there is no longer a single solution and players must negotiate ad-hoc both what they want to solve and how to actually solve it. Furthermore, we don't just play games we also create them. Consider the game of Calvinball which has only one rule: "you make up the rules as you go." These are the kinds of ad-hoc games a pair of children might play in the backseat of a car on a long road trip, but they characterize everyday interactions such as politeness, office politics or even the rules of engagement just to name a few examples.

Human social interactions are negotiated in the moment, using norms and morals learned throughout development, by exploiting abstract cognitive theories of other agents. This thesis is a step towards understanding these cognitive abilities from the perspective of reverse-engineering i.e., recreating them in mathematically precise models. These models and algorithms point to new ways of engineering social machines that understand, learn from, and cooperate with people and could narrow the gap between human and machine intelligence.

References

Adams, J. S. (1965). Inequity in social exchange. *Advances in experimental social psychology*, *2*(267-299).

Allport, G. W. (1954). *The nature of prejudice.* Addison-Wesley.

Arrow, K. J. (1965). *Aspects of the theory of risk-bearing*. Yrjö Jahnssonin Säätiö.

Axelrod, R. (1985). *The evolution of cooperation.* Basic Books.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396.

Ayars, A., & Nichols, S. (2017). Moral empiricism and the bias for act-based rules. *Cognition*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baker, M. C. (2002). *The atoms of language: The mind's hidden rules of grammar*. Basic books.

Bandura, A., & McDonald, F. J. (1963). Influence of social reinforcement and the behavior of models in shaping children's moral judgment. *The Journal of Abnormal and Social Psychology*, *67*(3), 274.

Baron, J., & Leshner, S. (2000). How serious are expressions of protected values? *Journal of Experimental Psychology: Applied*, *6*(3), 183.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational behavior and human decision processes*, *70*(1), 1–16.

Barragan, R. C., & Dweck, C. S. (2014). Rethinking natural altruism: Simple reciprocal interactions trigger children's benevolence. *Proceedings of the National Academy of Sciences*, *111*(48), 17071–17074.

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological bulletin*, *91*(1), 3.

Baunach, D. M. (2011). Decomposing trends in attitudes toward gay marriage, 1988–2006*. *Social Science Quarterly*, *92*(2), 346–363.

Baunach, D. M. (2012). Changing same-sex marriage attitudes in america from 1988 through 2010. *Public Opinion Quarterly*, *76*(2), 364–378.

Beal, M. J., Ghahramani, Z., & Rasmussen, C. E. (2002). The infinite hidden markov model. *Advances in neural information processing systems*, *1*, 577–584.

Bénabou, R., & Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, *126*(2), 805–855.

Bennett, J. (1974). The conscience of huckleberry finn. *Philosophy*, *49*(188), 123–134.

Binmore, K. G. (1994). *Game theory and the social contract: just playing* (Vol. 2). MIT press.

Binmore, K. G. (1998). *Game theory and the social contract: just playing* (Vol. 2). Mit Press.

Blake, P., McAuliffe, K., Corbit, J., Callaghan, T., Barry, O., Bowie, A., ... others (2015). The ontogeny of fairness in seven societies. *Nature*.

Blake, P. R., & McAuliffe, K. (2011). âĂIJi had so much it didnâĂŹt seem fairâĂİ: Eight-year-olds reject two forms of inequity. *Cognition*, *120*(2), 215–224.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Blonski, M., Ockenfels, P., & Spagnolo, G. (2011). Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence. *American Economic Journal: Microeconomics*, *3*(3), 164–192.

Bloom, P. (2010). How do morals change? *Nature*, *464*(7288), 490–490.

Bó, P. D. (2005). Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American Economic Review*, *95*(5), 1591–1604.

Bó, P. D., & Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *The American Economic Review*, *101*(1), 411–429.

Bodner, R., & Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions*, *1*, 105–26.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. OUP Oxford.

Bowling, M., Burch, N., Johanson, M., & Tammelin, O. (2015). Heads-up limit holdâĂŹem poker is solved. *Science*, *347*(6218), 145–149.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago press.

Bratman, M. (1987). *Intention, plans, and practical reason*.

Bratman, M. E. (1993). Shared intention. *Ethics*, 97–113.

Bratman, M. E. (2014). *Shared agency: A planning theory of acting together*. Oxford University Press.

Breitmoser, Y. (2015). Cooperation, but no reciprocity: Individual strategies in the repeated prisoner's dilemma. *American Economic Review*, *105*(9), 2882-2910.

Brewer, M. B., & Kramer, R. M. (1985). The psychology of intergroup attitudes and behavior. *Annual review of psychology*, *36*(1), 219–243.

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *352*(6282), 220–224.

Brown, N., & Sandholm, T. (2017). Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, eaao1733.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 861–898.

Campbell, M., Hoane Jr, A. J., & Hsu, F.-h. (2002). Deep blue. *Artificial intelligence*, *134*(1-2), 57–83.

Campbell, R. (2014). Reflective equilibrium and moral consistency reasoning. *Australasian Journal of Philosophy*, *92*(3), 433–451.

Campbell, R., & Kumar, V. (2012). Moral reasoning on the ground*. *Ethics*, *122*(2), 273–312.

Carmel, D., & Markovitch, S. (1996). Learning models of intelligent agents. In *Aaai/iaai, vol. 1* (pp. 62–67).

Chater, N., Reali, F., & Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences*, *106*(4), 1015–1020.

Chisholm, R. M. (1966). Freedom and action. In K. Lehrer (Ed.), *Freedom and determinism.* New York, NY: Random House.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*, 93–115.

Choi, Y.-j., & Luo, Y. (2015). 13-month-oldsâĂŹ understanding of social interactions. *Psychological Science*, *26*(3), 274–283.

Chomsky, N. (1980). *Rules and representations.* Blackwell.

Chomsky, N. (1981). *Lectures on government and binding: The pisa lectures.* Walter de Gruyter.

Choshen-Hillel, S., Shaw, A., & Caruso, E. M. (2015). Waste management: How reducing partiality can promote efficient resource allocation. *Journal of personality and social psychology*, *109*(2), 210.

Choshen-Hillel, S., & Yaniv, I. (2011). Agency and the construction of social preference: Between inequality aversion and prosocial behavior. *Journal of personality and social psychology*, *101*(6), 1253.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in cognitive sciences*, *7*(7), 300–307.

Coase, R. H. (1960). The problem of social cost. *The journal of Law and Economics*, *56*(4), 837–877.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial intelligence*, *42*(2), 213–261.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture*, *163*, 163–228.

Costa-Gomes, M., Crawford, V. P., & Broseta, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 1193–1235.

Cowan, P. A., Longer, J., Heavenrich, J., & Nathanson, M. (1969). Social learning and piaget's cognitive theory of moral development. *Journal of Personality and Social Psychology*, *11*(3), 261.

Crisp, R. J., & Turner, R. N. (2009). Can imagined interactions produce positive perceptions?: Reducing prejudice through simulated social contact. *American Psychologist*, *64*(4), 231.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363–366.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273–292.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80.

DeBruine, L. M. (2002). Facial resemblance enhances trust. *Proceedings of the Royal Society of London B: Biological Sciences*, *269*(1498), 1307–1312.

De Cote, E. M., & Littman, M. L. (2008). A polynomial-time nash equilibrium algorithm for repeated stochastic games. In *24th conference on uncertainty in artificial intelligence.*

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, *108*(32), 13335–13340.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

DeScioli, P. (2016). The side-taking hypothesis for moral judgment. *Current Opinion in Psychology*, *7*, 23–27.

Deutsch, D. (2011). *The beginning of infinity: Explanations that transform the world*. Penguin UK.

Diuk, C., Cohen, A., & Littman, M. L. (2008). An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on machine learning* (pp. 240–247).

Doshi-Velez, F., Wingate, D., Roy, N., & Tenenbaum, J. B. (2010). Nonparametric bayesian policy priors for reinforcement learning. In *Advances in neural information processing systems* (pp. 532–540).

Dungan, J., Waytz, A., & Young, L. (2014). Corruption in the context of moral trade-offs. *Journal of Interdisciplinary Economics*, *26*(1-2), 97–118.

Evans, O., Stuhlmüller, A., & Goodman, N. D. (2016). Learning the preferences of ignorant, inconsistent agents. In *Aaai* (pp. 323–329).

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. (2004). *Reasoning about knowledge*. MIT press.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairnessâĂŤintentions matter. *Games and Economic Behavior*, *62*(1), 287–303.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293–315.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817–868.

Feiman, R., Carey, S., & Cushman, F. (2015). InfantsâĂŹ representations of othersâĂŹ goals: Representing approach over avoidance. *Cognition*, *136*, 204–214.

Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.

Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, *99*(4), 689.

Friedman, D. D. (2001). *Law's order: what economics has to do with law and why it matters*. Princeton University Press.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual review of psychology*, *63*, 287–313.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games* (Vol. 2). MIT press.

Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, *54*(3), 533–554.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., ... others (2002). The structure of haplotype blocks in the human genome. *Science*, *296*(5576), 2225–2229.

Gal, Y., & Pfeffer, A. (2008). Networks of influence diagrams: A formalism for representing agents' beliefs and decision-making processes. *Journal of Artificial Intelligence Research*, *33*(1), 109–147.

Galinsky, A., & Schweitzer, M. (2015). *Friend and foe: When to cooperate, when to compete, and how to succeed at both*. Random House.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 37th annual conference of the cognitive science society.*

Gintis, H. (2009). *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton University Press.

Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multiagent settings. *J. Artif. Intell. Res.(JAIR)*, *24*, 49–79.

Gmytrasiewicz, P. J., & Durfee, E. H. (2000). Rational coordination in multi-agent environments. *Autonomous Agents and Multi-Agent Systems*, *3*(4), 319–350.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Lawrence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–653). MIT Press.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological review*, *118*(1), 110.

Govrin, A. (2014). The abc of moral development: an attachment approach to moral judgment. *Frontiers in psychology*, *5*, 6.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Graham, J., Meindl, P., Beall, E., Johnson, K. M., & Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, *8*, 125–130.

Gray, K., Rand, D. G., Ert, E., Lewis, K., Hershman, S., & Norton, M. I. (2014). The emergence of "us and them" in 80 lines of code modeling group genesis in homogeneous populations. *Psychological science*, 0956797614521816.

Greene, J. (2014). *Moral tribes: emotion, reason and the gap between us and them*. Atlantic Books Ltd.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Grosz, B. J., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, *86*(2), 269–357.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 3909–3917). Curran Associates, Inc. Retrieved from `http://papers.nips.cc/paper/6420-cooperative-inverse-reinforcement-learning.pdf`

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, *56*(4), 843–887.

Hamann, K., Warneken, F., Greenberg, J. R., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, *476*(7360), 328–331.

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers evidence for an innate moral core. *Current Directions in Psychological Science*, *22*(3), 186–193.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me= bad infants prefer those who harm dissimilar others. *Psychological science*, *24*(4), 589–594.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*(3859), 1243–1248.

Heider, F. (1958). *The psychology of interpersonal relations*. Psychology Press.

Hein, G., Engelmann, J. B., Vollberg, M. C., & Tobler, P. N. (2016). How learning shapes the empathic brain. *Proceedings of the National Academy of Sciences*, *113*(1), 80–85.

Hein, G., Morishima, Y., Leiberg, S., Sul, S., & Fehr, E. (2016). The brain's functional network architecture reveals human motives. *Science*, *351*(6277), 1074–1078.

Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, *91*(2), 73–78.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and human behavior*, *22*(3), 165–196.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *science*, *317*(5843), 1360–1366.

Heyes, C. (2016). Who knows? metacognitive social learning strategies. *Trends in cognitive sciences*.

Hoffman, M. L. (1975). Altruistic behavior and the parent-child relationship. *Journal of personality and social psychology*, *31*(5), 937.

Hoffman, M. L. (2001). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

Hook, J., & Cook, T. D. (1979). Equity theory and the cognitive ability of children. *Psychological Bulletin*, *86*(3), 429.

Horne, Z., Powell, D., & Hummel, J. (2015). A single counterexample leads to moral belief revision. *Cognitive science*, *39*(8), 1950–1964.

House, B. R., Silk, J. B., Henrich, J., Barrett, H. C., Scelza, B. A., Boyette, A. H., ... Laurence, S. (2013). Ontogeny of prosocial behavior across diverse societies. *Proceedings of the National Academy of Sciences*, *110*(36), 14586–14591.

Hume, D. (1738). *A treatise of human nature*.

Humphrey, N. K. (1976). The social function of intellect. In *Growing points in ethology* (pp. 303–317). Cambridge University Press.

Hurka, T. (2003). *Virtue, vice, and value*.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38.

Kalai, E., & Lehrer, E. (1993). Rational learning leads to nash equilibrium. *Econometrica: Journal of the Econometric Society*, *61*(5), 1019–1045.

Keasey, C. B. (1973). Experimentally induced changes in moral opinions and reasoning. *Journal of Personality and Social Psychology*, *26*(1), 30.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–1243.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, *10*(3), 307–321.

Kempe, A. (1997). Finite state transducers approximating hidden markov models. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics* (pp. 460–467).

Kiley Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental science*, *16*(2), 209–226.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th annual conference of the cognitive science society.*

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *Proceedings of the 38th annual conference of the cognitive science society.*

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In *Proceedings of the 39th annual conference of the cognitive science society.*

Kleiman-Weiner, M., Tenenbaum, J. B., & Zhou, P. (in press). Non-parametric bayesian inference of strategies in infinitely repeated games. *Econometrics Journal*.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190–194.

Kohlberg, L. (1981). *The philosophy of moral development: moral stages and the idea of justice*. Harper & Row.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Krienen, F. M., Tu, P.-C., & Buckner, R. L. (2010). Clan mentality: evidence that the medial prefrontal cortex responds to close others. *The Journal of Neuroscience*, *30*(41), 13906–13915.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.

Lake, B., & Tenenbaum, J. (2010). Discovering structure by learning sparse graph. In *Proceedings of the 33rd annual cognitive science conference.*

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Levesque, H. J., Cohen, P. R., & Nunes, J. H. (1990). On acting together. In *Aaai* (Vol. 90, pp. 94–99).

Liddle, B., & Nettle, D. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology*, *4*(3-4), 231–244.

Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature*, *445*(7129), 727–731.

Littman, M. L., & Stone, P. (2005). A polynomial-time nash equilibrium algorithm for repeated games. *Decision Support Systems*, *39*(1), 55–66.

Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley.

Magid, R. W., & Schulz, L. E. (2017). Moral alchemy: How love changes norms. *Cognition*.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121.

Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. *Intentions and intentionality: Foundations of social cognition*, 45–67.

Malle, B. F., Moses, L. J., & Baldwin, D. A. (2001). *Intentions and intentionality: Foundations of social cognition*. MIT press.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company*, *1*(2).

McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., & Gureckis, T. (2012). psiturk.

Mikhail, J. (2006). The poverty of the moral stimulus.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences*, *11*(4), 143–152.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in cognitive sciences*.

Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, *16*(1), 69–88.

Monin, B. (2007). Holier than me? threatening social comparison in the moral domain. *Revue internationale de psychologie sociale*, *20*(1), 53–68.

Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: resenting those who do the right thing. *Journal of personality and social psychology*, *95*(1), 76.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*.

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

Nagel, T. (1989). *The view from nowhere*. Oxford University Press.

Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H.-Y. (2016). Rational learners and moral rules. *Mind & Language*, *31*(5), 530–554.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530–542.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. MIT press Cambridge, MA:.

Nook, E. C., Ong, D. C., Morelli, S. A., Mitchell, J. P., & Zaki, J. (2016). Prosocial conformity prosocial norms generalize across behavior and empathy. *Personality and Social Psychology Bulletin*, 0146167216649932.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, *364*(6432), 56.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563.

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogenous populations. *Nature*, *355*(6357), 250.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298.

Ostron, E. (1990). *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? a review and assessment of research and practice. *Annual review of psychology*, *60*, 339–367.

Panella, A., & Gmytrasiewicz, P. J. (2015). Nonparametric bayesian learning of other agents? policies in interactive pomdps. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems* (pp. 1875–1876).

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, *90*(5), 751.

Peysakhovich, A., & Rand, D. G. (2015). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, *62*(3), 631–647.

Pineau, J., Gordon, G., Thrun, S., et al. (2003). Point-based value iteration: An anytime algorithm for pomdps. In *Ijcai* (Vol. 3, pp. 1025–1032).

Pinker, S. (1997). *How the mind works*. Norton.

Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Penguin.

Pizarro, D. (2000). Nothing more than feelings? the role of emotions in moral judgment. *Journal for the Theory of Social Behaviour*, *30*(4), 355–375.

Pizarro, D. A., Detweiler-Bedell, B., & Bloom, P. (2006). The creativity of everyday moral reasoning. *Creativity and reason in cognitive development*, 81–98.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. *The social psychology of morality: Exploring the causes of good and evil*, 91–108.

Platt, R., Tedrake, R., Kaelbling, L., & Lozano-Perez, T. (2010, June). Belief space planning assuming maximum likelihood observations. In *Proceedings of robotics: Science and systems*. Zaragoza, Spain.

Popper, K. S. (2012). *The open society and its enemies*. Routledge.

Posner, R. A. (1973). *Economic analysis of law*. Little Brown and Company.

Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, *110*(41), E3965–E3972.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*(1/2), 122–136.

Prince, A., & Smolensky, P. (2008). *Optimality theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Prinz, J. (2007). *The emotional construction of morals*. Oxford University Press.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, *118*(1), 57.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science*, *325*(5945), 1272–1275.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in cognitive sciences*, *17*(8), 413.

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., ... Laland, K. N. (2010). Why copy others? insights from the social learning strategies tournament. *Science*, *328*(5975), 208–213.

Rendell, L., Fogarty, L., Hoppitt, W. J., Morgan, T. J., Webster, M. M., & Laland, K. N. (2011). Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends in cognitive sciences*, *15*(2), 68–76.

Rhodes, M. (2012). Naïve theories of social groups. *Child development*, *83*(6), 1900–1916.

Rhodes, M., & Chalik, L. (2013). Social categories as markers of intrinsic interpersonal obligations. *Psychological science*, *24*(6), 999–1006.

Rhodes, M., & Wellman, H. (2016). Moral learning as intuitive theory revision. *Cognition*.

Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, *35*(2), 395–405.

Rubinstein, A. (1986). Finite automata play the repeated prisoner's dilemma. *Journal of economic theory*, *39*(1), 83–96.

Scanlon, T. M. (1975). Preference and urgency. *The Journal of Philosophy*, *72*(19), 655–669.

Scanlon, T. M. (2009). *Moral dimensions*. Harvard University Press.

Schäfer, M., Haun, D. B., & Tomasello, M. (2015). Fair is not fair everywhere. *Psychological science*, 0956797615586188.

Schult, C. A. (2002). Children's understanding of the distinction between intentions and desires. *Child Development*, *73*(6), 1727–1747.

Searle, J. (1969). *Intentionality: An essay in the philosophy of mind.* New York, NY: Cambridge University Press.

Sen, A., & Hawthorn, G. (1988). *The standard of living*. Cambridge University Press.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Shachter, R. D. (1986). Evaluating influence diagrams. *Operations research*, *34*(6), 871–882.

Shaw, A. (2013). Beyond âĂIJto share or not to shareâĂİ the impartiality account of fairness. *Current Directions in Psychological Science*, *22*(5), 413–417.

Shaw, A., & Olson, K. (2014). Fairness as partiality aversion: The development of procedural justice. *Journal of Experimental Child Psychology*, *119*, 40–53.

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General*, *141*(2), 382.

Shoham, Y. (1993). Agent-oriented programming. *Artificial intelligence*, *60*(1), 51–92.

Shook, N. J., & Fazio, R. H. (2008). Interracial roommate relationships an experimental field test of the contact hypothesis. *Psychological Science*, *19*(7), 717–723.

Sigelman, C. K., & Waitzman, K. A. (1991). The development of distributive justice orientations: Contextual influences on children's resource allocations. *Child Development*, 1367–1378.

Sigmund, K. (2010). *The calculus of selfishness*. Princeton University Press.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... others (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

Singer, P. (1981). *The expanding circle*. Clarendon Press Oxford.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 70–82.

Sinnott-Armstrong, W., Mallon, R., Mccoy, T., & Hull, J. G. (2008). Intention, temporal order, and moral judgments. *Mind & Language*, *23*(1), 90–106.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2012). A causal model of intentionality judgment. *Mind & Language*, *27*(2), 154–180.

Smetana, J. G. (2006). Social-cognitive domain theory: Consistencies and variations in childrenâĂŹs moral and social judgments. *Handbook of moral development*, 119–

153.

Smith, A. (1759). *The theory of moral sentiments*.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial life*, *9*(4), 371–386.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of experimental social psychology*, *27*(1), 76–105.

Suchow, J. W., Bourgin, D. D., & Griffiths, T. L. (2017). Evolution in mind: Evolutionary dynamics, cognitive processes, and bayesian inference. *Trends in cognitive sciences*, *21*(7), 522–530.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social philosophy and policy*, *10*(01), 69–89.

Sugden, R. (2003). The logic of team reasoning. *Philosophical explorations*, *6*(3), 165–181.

Takagishi, H., Kameshima, S., Schug, J., Koizumi, M., & Yamagishi, T. (2010). Theory of mind enhances preference for fairness. *Journal of experimental child psychology*, *105*(1), 130–137.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, *101*(476), 1566–1581.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*(6022), 1279.

Tesauro, G. (1995). Td-gammon: A self-teaching backgammon program. In *Applications of neural networks* (pp. 267–285). Springer.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in cognitive sciences*, *7*(7), 320–324.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of personality and social psychology*, *78*(5), 853.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

Tomasello, M. (1999). *The cultural origins of human cognition.* Harvard University Press.

Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(05), 675–691.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly review of biology*, 35–57.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems* (pp. 1874–1882).

Van Gael, J., Saatci, Y., Teh, Y. W., & Ghahramani, Z. (2008). Beam sampling for the infinite hidden markov model. In *Proceedings of the 25th international conference on machine learning* (pp. 1088–1095).

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard university press.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. *The Oxford handbook of thinking and reasoning*, 364–389.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Watanabe, T., Takezawa, M., Nakawake, Y., Kunimatsu, A., Yamasue, H., Nakamura, M., … Masuda, N. (2014). Two distinct neural mechanisms underlying indirect reciprocity. *Proceedings of the National Academy of Sciences*, *111*(11), 3990–3995.

Wattles, J. (1997). *The golden rule*.

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of personality and social psychology*, *45*(5), 961.

Wellman, H. M. (1992). *The child's theory of mind.*

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, *131*(3410), 1355–1358.

Wood, F., Gasthaus, J., Archambeau, C., James, L., & Teh, Y. W. (2011). The sequence memoizer. *Communications of the ACM*, *54*(2), 91–98.

Wright, J. C., & Bartsch, K. (2008). Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly (1982-)*, 56–85.

Wright, S. C., Aron, A., McLaughlin-Volpe, T., & Ropp, S. A. (1997). The extended contact effect: Knowledge of cross-group friendships and prejudice. *Journal of Personality and Social psychology*, *73*(1), 73.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12).

Young, H. P. (2001). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240.

Zagorsky, B. M., Reiter, J. G., Chatterjee, K., & Nowak, M. A. (2013). Forgiver triumphs in alternating prisoner's dilemma. *PloS one*, *8*(12), e80814.