Précis for "Computational Foundations of Human Social Intelligence"

Max Kleiman-Weiner

Human social intelligence is uniquely powerful. 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 scale and sophistication unparalleled in other animal species (Vygotsky, 1978; Warneken & Tomasello, 2006; Herrmann, Call, Hernández-Lloreda, Hare, & Tomasello, 2007; Spelke & Kinzler, 2007; Hamlin, 2013). This thesis seeks to understand these everyday feats of social intelligence in computational terms. What cognitive representations and processes underlie these abilities and what are their origins? How can we apply these cognitive principles to build social machines that can understand, learn from, and cooperate with people?

While cooperation is essential and beneficial, it 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 collective benefits. This can lead to a "tragedy of the commons" where cooperation is not self-sustaining (Hardin, 1968; Trivers, 1971). In addition to the challenge of incentives, 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 commonsense social skills of a kindergartner.

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.

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.

Throughout this thesis I propose 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; Baker, Saxe, & Tenenbaum, 2009; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). They also allow us to predict what another 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.

Evolution & Abstract Reciprocity

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 on their general applicability.

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 with a fixed set of labeled actions. 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 across which 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 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 evolutionary 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 and partners.

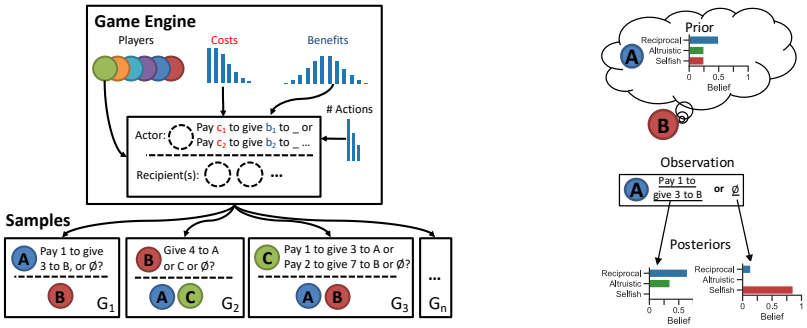

Figure 1. (left) Game engine which creates an infinitude of unique social choices where the number of players, the number of actions available to the decision making player, and the costs and/or benefits to each affected player are sampled from stochastic distributions. Actions are not labeled so decisions and inference must be made in terms of cost/benefit analysis. (right) Recursive Bayesian inference produces rational belief updates after observing A either pay a cost to help B or observing A withhold help.

I show that such a strategy – specifically, a conditional cooperator that uses recursive 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 (Figure 1). 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, we see that cognitive complexity enables the evolution of cooperation more effectively than purely behaviorist strategies. These results are also suggestive of how the challenge of cooperation can drive the evolution of cognitive complexity – a defining feature of humankind.

Development and Moral Learning

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?

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 (Kleiman-Weiner, Saxe, & Tenenbaum, 2017) . I 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. The 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.

In-the-Moment Social Cognition

Learning to Cooperate and Compete

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 (Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016). These games can be played intuitively by people (Figure 2).

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

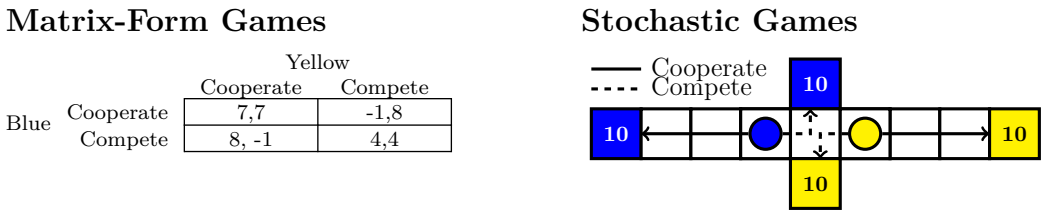

Figure 2. (left) A social dilemma written as a matrix-form game. If both agents choose cooperate they will collectively be better well off than if they both choose compete. However in any single interaction, either agent would be materially better off by choosing to compete. (right) A social dilemma as a two-player stochastic game. Agents (circles) score points by moving to squares of their own color. The arrows show example plans that realize either cooperative and competitive outcomes.

reciprocate the inferred intention. In new environments, people generalize abstract intentions like cooperation and competition by executing a novel 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 or "view from nowhere" (Nagel, 1986). 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 (Kleiman-Weiner, Tenenbaum, & Zhou, in press). 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.

Reputation and fairness

In **Chapter 6** I study how humans allocate the spoils of a cooperative endeavor. The ability to flexibly allocate a joint reward expands 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 (Kleiman-Weiner, Shaw, & Tenenbaum, 2017). 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.

Intention inference in moral judgment

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 (Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Halpern & Kleiman-Weiner, 2018). 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.

Conclusion

I have shown that human social interactions are negotiated in the moment using abstract causal theories of other agents, guided by norms and morals learned throughout development, which have been shaped by the evolutionary challenges of cooperation. This thesis is a step towards understanding these cognitive abilities from the perspective of reverse-engineering i.e., recreating these abilities in mathematically precise models. 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. Together, these models shine light on how the scale and scope of human social behavior is ultimately grounded in the sophistication of our social intelligence.

References

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

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.

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

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

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

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

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

Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Aaai*.

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.

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

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

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.

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.

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*.

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

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

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

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.

Ostron, E. (1990). *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press.

Pinker, S. (1997). *How the mind works*. Norton.

Prinz, J. (2007). *The emotional construction of morals*. Oxford University Press.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in cognitive sciences*, *17*(8), 413.

Sigmund, K. (2010). *The calculus of selfishness*. Princeton University Press.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

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.

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.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard university press.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Wellman, H. M. (1992). *The child's theory of mind.*