Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction

Max Kleiman-Weiner[1] (maxkw@mit.edu), Mark K. Ho (mark_ho@brown.edu)[2], Joseph L. Austerweil[2] (joseph_austerweil@brown.edu), Michael L. Littman[3] (mlittman@cs.brown.edu), Joshua B. Tenenbaum[1] (jbt@mit.edu)

[1]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

[2]Cognitive, Linguistic, and Psychological Sciences / [3]Computer Science, Brown University, Providence, RI 02912

Abstract

Successfully navigating the social world requires reasoning about both high-level strategic goals, such as whether to cooperate or compete, as well as the low-level actions needed to achieve those goals. While previous work in experimental game theory has examined the former and work on multi-agent systems has examined the later, there has been little work investigating behavior in environments that require simultaneous planning and inference across both levels. We develop a hierarchical model of social agency that infers the intentions of other agents, strategically decides whether to cooperate or compete with them, and then executes either a cooperative or competitive planning program. Learning occurs across both high-level strategic decisions and low-level actions leading to the emergence of social norms. We test predictions of this model in multi-agent behavioral experiments using rich video-game like environments. By grounding strategic behavior in a formal model of planning, we develop abstract notions of both cooperation and competition and shed light on the computational nature of joint intentionality.

Keywords: joint intention, cooperation, coordination, reinforcement learning, teams

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 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 oth-

Matrix-Form Games

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

Figure 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.

Stochastic Games

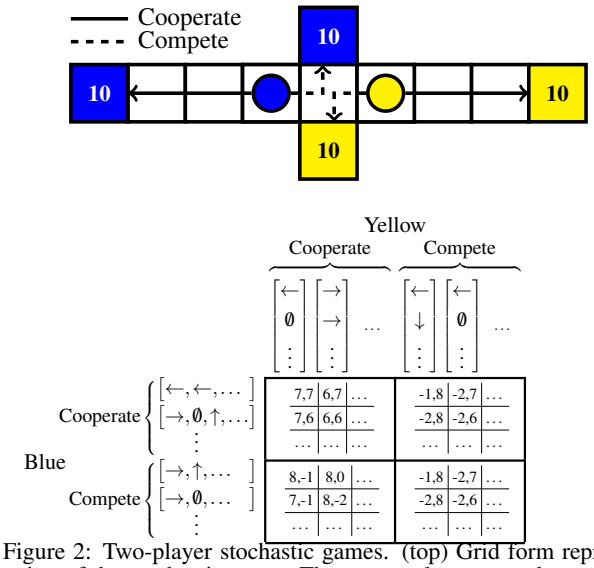

Figure 2: Two-player stochastic games. (top) Grid form representation of the stochastic game. The arrows show example strategies that can be used to realize both cooperative and competitive outcomes. (bottom) Matrix representation of the strategy space, with low-level strategies sorted by a high-level goal. The arrows correspond to moving in a specific direction and the $\emptyset$ 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.

ers 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, Carpenter, Call, Behne, & Moll, 2005).

Naturalistic Games

Game-theoretic investigations of social behavior often represent strategic interactions as *matrix-form* games like the one shown in Figure 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 2 shows an example of one of these multi-agent planning environments that is conceptually related to the social dilemma shown in Figure 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.

Model

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.

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). 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:

$$P(a_1,a_2|s) = \pi^G(s,a_1,a_2) \propto e^{(\beta Q^G(s,a_1,a_2))}$$

$$Q^G(s,a_1,a_2) = \sum_{s'} P(s'|s,a_1,a_2)[U^G(s',s,a_1,a_2) + \gamma \max_{(a_1',a_2')} Q^G(s',a_1',a_2')]$$

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 (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 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, Melkonyan, Zeitoun, & Chater, 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 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.

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 (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 (Camerer et al., 2004; Yoshida, Dolan, & Friston, 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.

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).

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

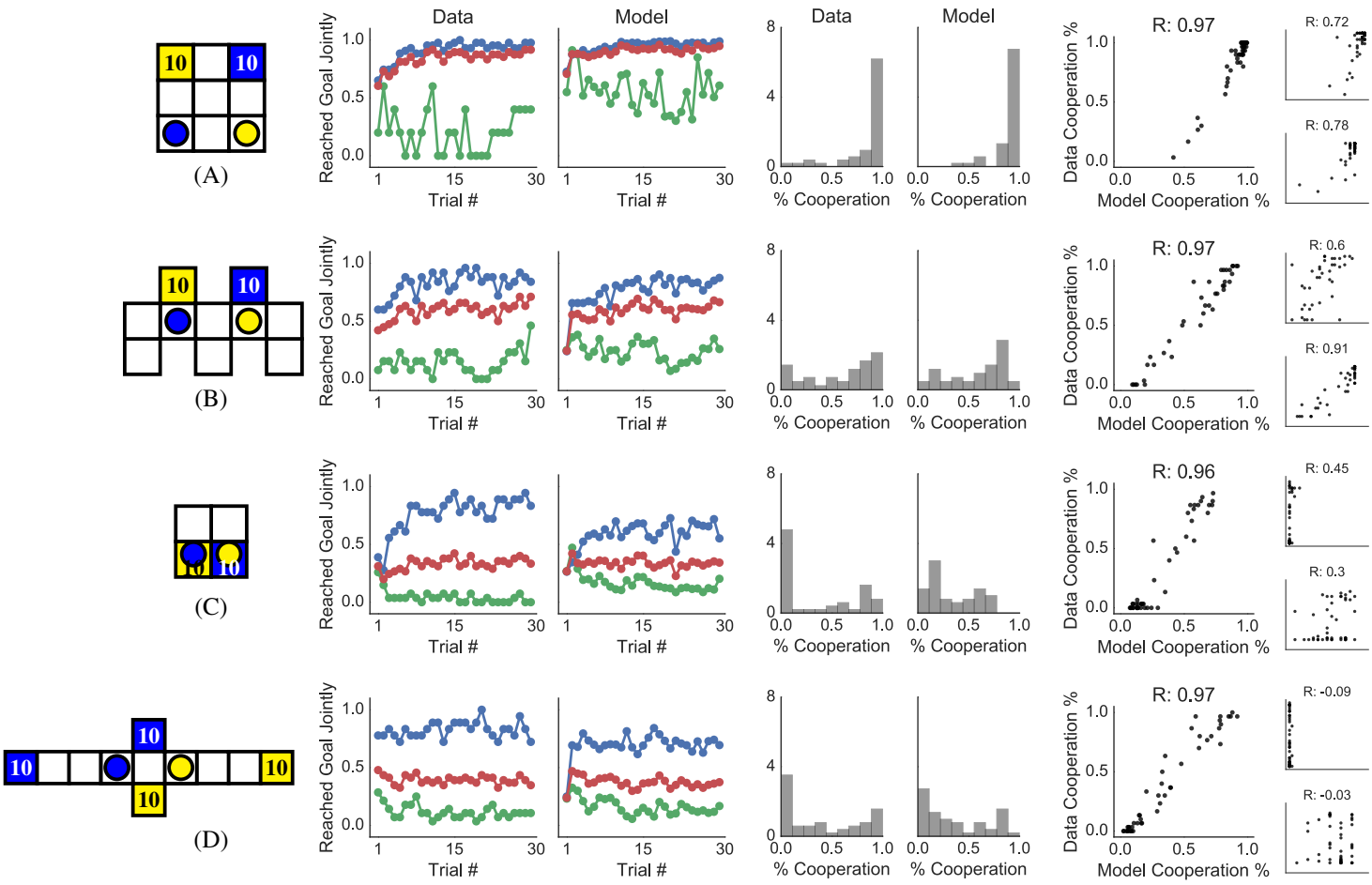

Figure 3: Participant data and model predictions for four environments. Each row shows data and model predictions for the environment in column 1 which was repeated 30 times. Rows 1 and 2 are coordination games and rows 3 and 4 are social dilemmas. Column 2 shows the average rate of cooperation for each 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.

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 model parameters were used for all pairs of participants.

Figure 3 shows the results of the behavioral experiments and the model predictions for four environments ($\approx$ 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 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 qualitatively 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.

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 attribute cooperative and competitive intentions to other agents.

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. 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).

Acknowledgement This work was supported by a Hertz Foundation Fellowship, NSF-GRFP, the Center for Brains, Minds and Machines (CBMM), NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333. We thank Alejandro Vientós, Banti Gheneti, and Paul Masterson.

References

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

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

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

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

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.*

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

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.

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

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

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.

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

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

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

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

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.

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.

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

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12).

Cognition xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Learning a commonsense moral theory

Max Kleiman-Weiner *, Rebecca Saxe, Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, United States

ARTICLE INFO

Article history:
Received 1 April 2016
Revised 26 February 2017
Accepted 9 March 2017
Available online xxxx

Keywords:
Moral learning
Hierarchical Bayesian models
Social cognition
Moral change
Value alignment

ABSTRACT

We introduce 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. Introduction

> *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 (1989), 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 (Blake et al., 2015; Henrich et al., 2001; 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?

* Corresponding author at: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave, 46-4053, Cambridge, MA 02139, United States.
E-mail address: maxkw@mit.edu (M. Kleiman-Weiner).

http://dx.doi.org/10.1016/j.cognition.2017.03.005
0010-0277/

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 (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 (Pinker, 2011; Singer, 1981). 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', Popper, 2012; Wattles, 1997); 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 (Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Ullman et al., 2009). 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 (Barragan & Dweck, 2014; Hamlin, 2013; Hamlin, Mahajan, Liberman, & Wynn, 2013; Kohlberg, 1981; Powell & Spelke, 2013; Rhodes, 2012; Rhodes & Chalik, 2013; Rhodes & Wellman, 2016; 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; House et al., 2013; Sigelman & Waitzman, 1991). Different cultures or subcultures might weight these principles in different ways, generating different moral theories (Graham, Meindl, Beall, Johnson, & Zhang, 2016; Schäfer, Haun, & Tomasello, 2015) 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.

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

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, Kemp, Griffiths, & Goodman, 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; Mikhail, 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" (Baker, 2002; Chomsky, 1981) 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 (Ayars & Nichols, 2017; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Nichols, Kumar, Lopez, Ayars, & Chan, 2016; Perfors, Tenenbaum, & Regier, 2011; Tenenbaum et al., 2011; Xu & Tenenbaum, 2007), 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. (Baker, Saxe, & Tenenbaum, 2009; Goodman, Tenenbaum, & Gerstenberg, 2015; Heider, 1958; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Malle, Moses, & Baldwin, 2001; Ullman et al., 2009). 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 (Chater, Reali, & Christiansen, 2009; Christiansen & Kirby, 2003; Griffiths & Kalish, 2007; Kirby, Cornish, & Smith, 2008; Niyogi, 2006; Smith, Kirby, & Brighton, 2003). 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, this issue). 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 (Delton, Krasnow, Cosmides, & Tooby, 2011; Henrich & Gil-White, 2001; Nowak, 2006; Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009; Rand & Nowak, 2013; Richerson & Boyd, 2008; Trivers, 1971). 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 (Pinker, 2011; Singer, 1981). 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 (Campbell & Kumar, 2012). 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 (Hoffman, 2001; Pizarro, 2000), imagination and stories (Bloom, 2010), analogical reasoning (Campbell & Kumar, 2012; Keasey, 1973), love, or involved contact (even imagined or vicarious) (Allport, 1954; Crisp & Turner, 2009; Paluck & Green, 2009; Pettigrew & Tropp, 2006; Shook & Fazio, 2008; Wright, Aron, McLaughlin-Volpe, & Ropp, 1997). 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.

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

2. 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 A 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:

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

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 $\gamma_i \geqslant 0.5$ they value others more than themselves. Generally speaking, γ_i determines the degree to which agent i is prosocial. Each $\alpha_{ij} \in [0,1]$ is the weight agent i places on the utility of agent j. Depending on the relative value of each α_{ij}, an agent acting under a level-1 moral theory will value some agents more than others. If $\alpha_{ij} > \alpha_{ik}$ 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 (Camerer, 2003). We consider further extensions to these representations in Appendix B.

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:

$$EU(a,s) = \sum_{s'} U(s')P(s'|a,s) \tag{2}$$

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):

$$P(a|s) = \frac{\exp(\beta EU(a,s))}{\sum_{a'\in A_s} \exp(\beta EU(a',s))} \tag{3}$$

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 α_{ij} define how each agent values the others. However, each α_{ij} 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 α_{ij} 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 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 (Krienen, Tu, & Buckner, 2010; Rilling et al., 2002; 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

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

under principle p and 0 otherwise. Specifically, $f^{\text{kin}}(i,j)=1$ if i and j are kin, $f^{\text{group}}(i,j)=1$ if i and j are in the same in-group and $f^{\text{all}}(i,j)=1$ for all $i \neq j$. $f^{\text{self}}(i,j)=1$ for all $i=j$. The $f^{\text{d-recip}}(i,j)=1$ if i and j have a reciprocal relationship and $f^{\text{i-recip}}(i,j)=1$ if both i and j are in the cooperative group (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.

Fig. 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 (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; 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 Eq. (1) as a function of weights over moral principles:

$$\alpha_{i,j}(W_i) = \phi_{i,j} + \sum_{p \in P} w_i^p \cdot f^p(i,j) \tag{4}$$

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 Eq. (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. Fig. 1c shows the $\alpha_{i,j}$ derived from the weights and relations of Fig. 1.

3. 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" (Goodman, Ullman, & Tenenbaum, 2011; Kemp et al., 2007; Kemp, Goodman, & Tenenbaum, 2010). 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.

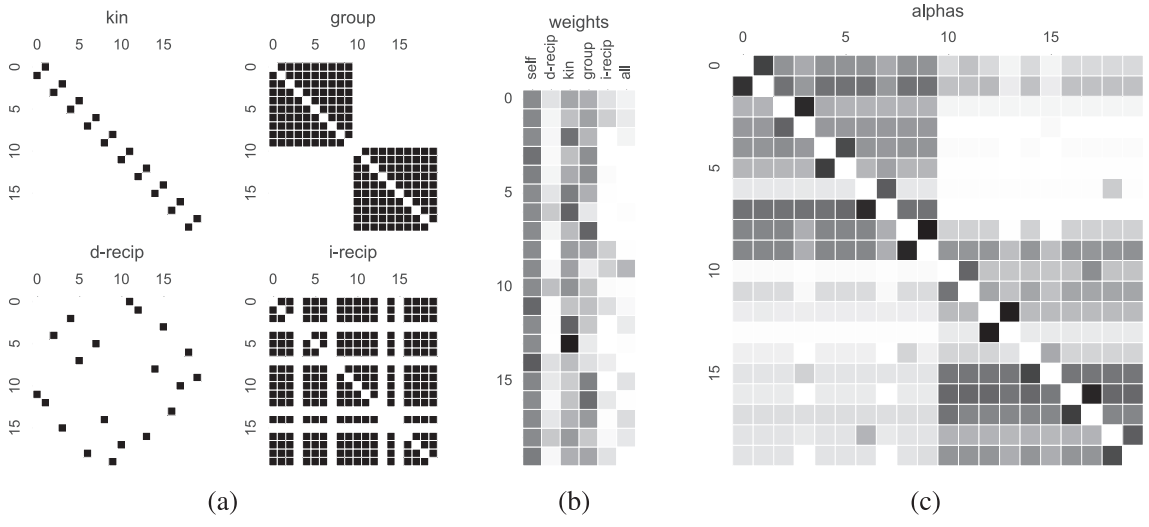

Fig. 1. A population of 20 agents used throughout this work. (a) Black squares indicate the presence of a relationship for each of the four principles shown. (b) The relative weights on each of the six principles for all 20 agents where each row is the weighting of principles of a single agent. Darker 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.

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

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:

$$P(W_i|(a_i^0,s^0),\ldots,(a_i^T,s^T)) \propto \sum_{f^{\text{d-recip}}} \sum_{f^{\text{i-recip}}} P(a_i^0,\ldots,a_i^T|s^0,\ldots,s^T,W_i, f^{\text{d-recip}},f^{\text{i-recip}})P(W_i)P(f^{\text{i-recip}})P(f^{\text{d-recip}})$$

where the likelihood $P(a_i^0,\ldots,a_i^T|s^0,\ldots,s^T,W_i,f^{\text{d-recip}},f^{\text{i-recip}})$ is probabilistic rational action as shown in Eq. (3) with the α_{ij} set by the weights over moral principles as shown in Eq. (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 (Lake & Tenenbaum, 2010):

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

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":

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

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:

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

where W^g_{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 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 Fig. 1b. Importantly, a learner does not know the W^g_{norm} for each group g in advance. The group average W^g_{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 Fig. 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_{norm} for each group, the learner could infer a single W_{norm} for all agents which would imply that the learner assumes moral theories do not systematically vary across groups. Furthermore, the W^g_{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

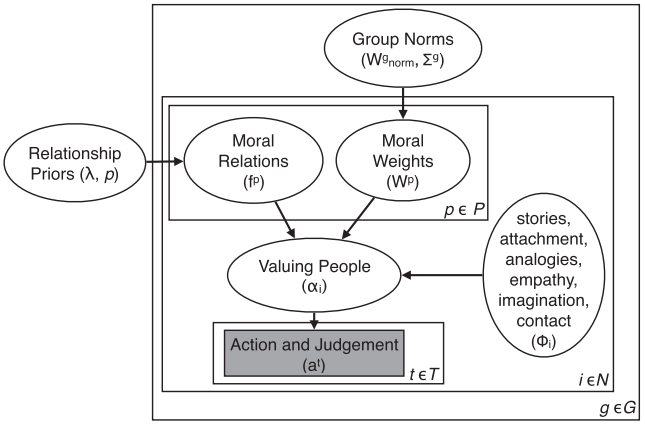

Fig. 2. Hierarchical probabilistic model for inferring latent moral theories from sparse behavior. T is the number of actions and judgments observed, N are the agents, P are moral principles and G are the groups. Actions and judgments are observed (shaded in gray).

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 A 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. Fig. 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 $P(f^{\text{d-recip}}, f^{\text{i-recip}}|(a_i^0, s^0), \ldots, (a_i^T, s^T))$. 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.

4. 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 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,

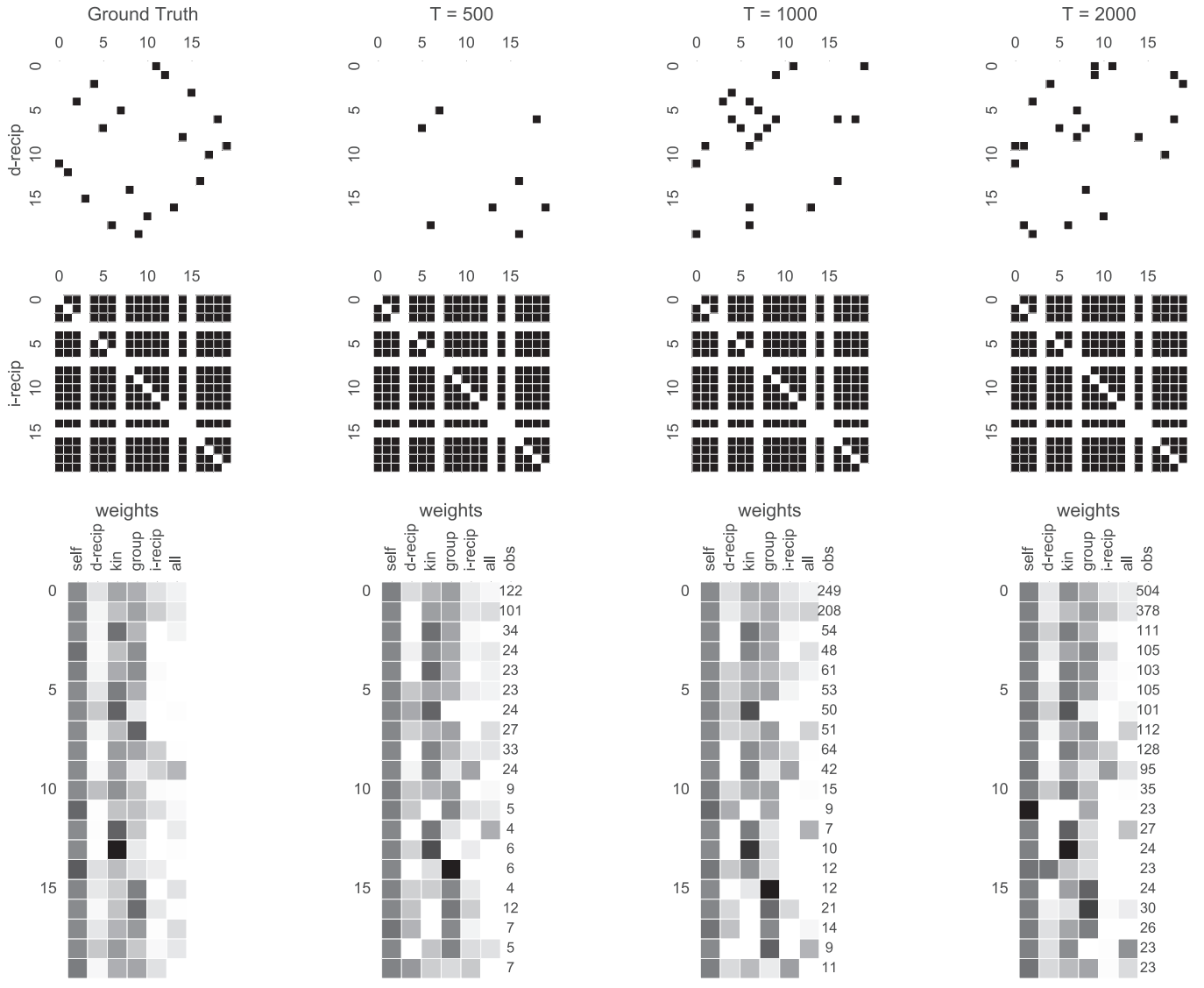

Fig. 3. Maximum a posteriori (MAP) estimate of beliefs from a learner observing behavior from the society shown in Fig. 1 under increasing observations ($T = \{500, 1000, 2000\}$). This learner is biased towards observing the behavior of agents 0 and 1. (top) Samples of the graph inference for the two 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 A for the details of the inference.

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

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.

4.1. 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 (Nowak, 2006; Rand & Nowak, 2013; Richerson & Boyd, 2008). 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 (Frith & Frith, 2012; Henrich & Gil-White, 2001; Heyes, 2016; Rendell et al., 2010; Rendell et al., 2011; Richerson & Boyd, 2008).

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, $\hat{w}_L$:

$$U_{\texttt{external}}(w_L|\hat{w}_L) = -\sum_{i\in N}\alpha_{L,i}(\hat{w}_L)\sum_{p\in P}(w_L^p - w_i^p)^2. \quad (6)$$

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 Eq. (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; Govrin, n.d.; Hoffman, 1975). 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 Eq. (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 $w_L^{\texttt{group}} > 0$ 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 (Fig. 1). Each of the 20 agents act as a caregiver (with a corresponding primitive attachment) to a single learner. Fig. 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 Fig. 4 shows, when this ϕ is high, the 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 Fig. 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.

4.2. 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, 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 (Campbell, 2014; Rawls, 1971). We argue that a

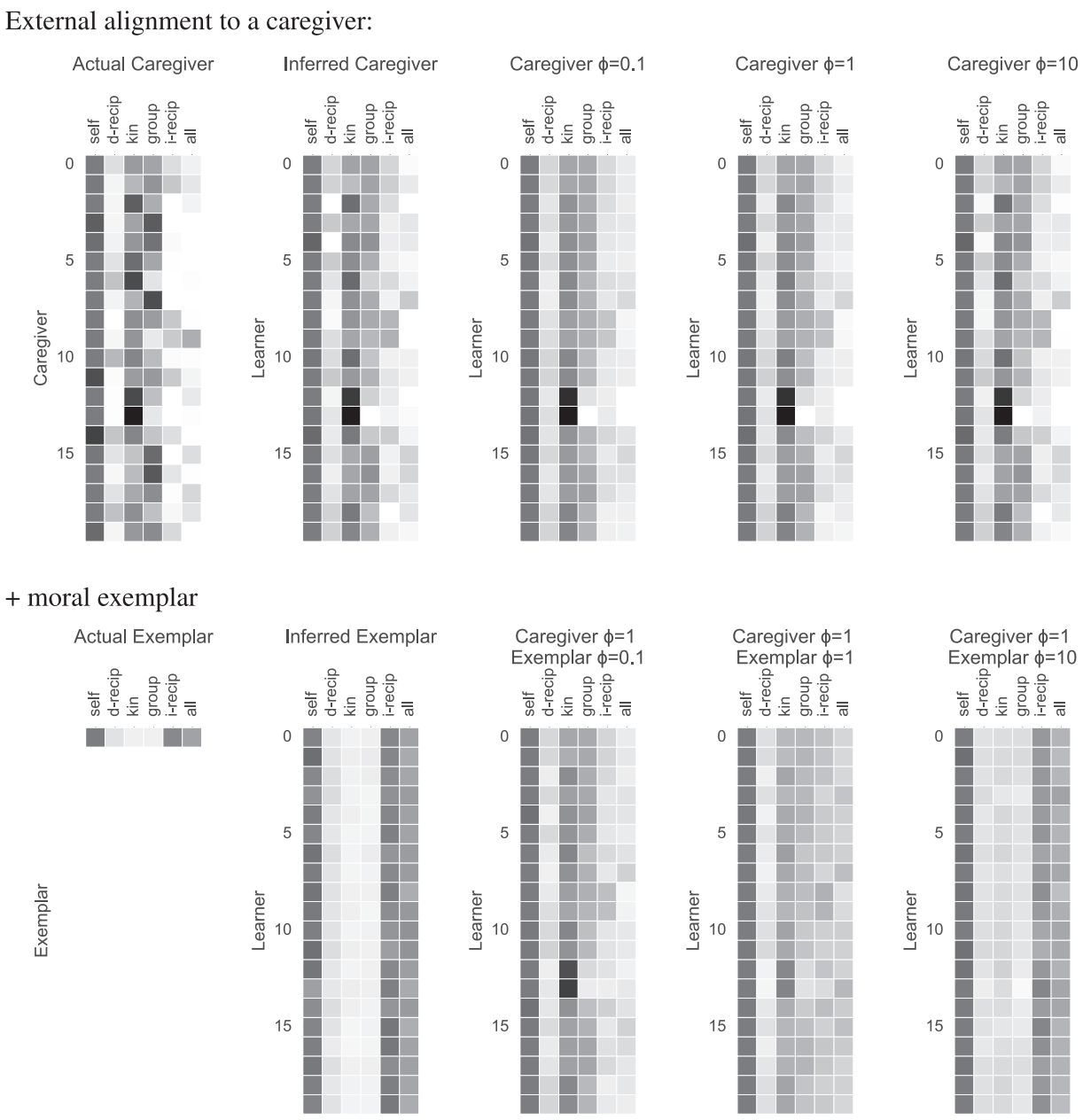

Fig. 4. External alignment with caregivers and moral exemplars. The "Actual" columns shows the actual weights for the caregivers of each of the 20 learners and the moral exemplar. The "Inferred" columns show the weights each learner infers about the weights over principles used by their own caregiver (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 Eq. (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.

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 (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 (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:

$$U_{\texttt{internal}}(w_L|\hat{w}_L) = -\sum_{i\in N}\left[\alpha_{L,i}(\hat{w}_L) - \sum_{p\in P} w_L^p \cdot f^p(L,i)\right]^2. \quad (7)$$

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 $\sum_{p\in P} w_L^p \cdot f^p(L,i)$ and the $\phi_{L,i}$ as shown in Eq. (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 Fig. 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 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. Fig. 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

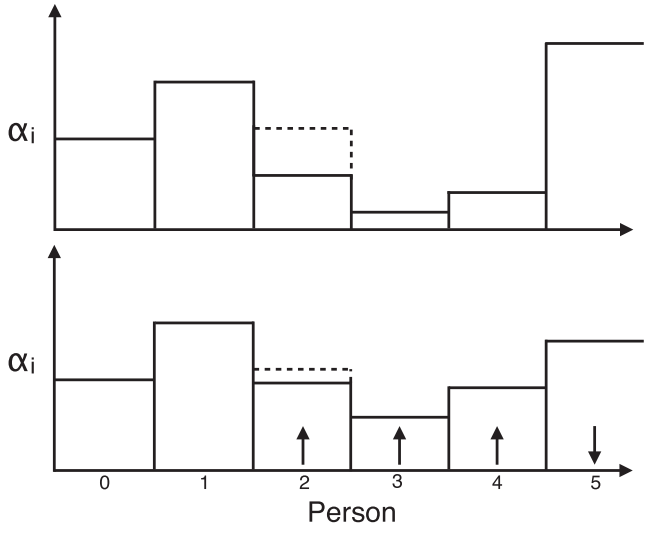

Fig. 5. Internal moral alignment through inconsistency reduction. (a, top) Schematic of a learner's current moral theory $\hat{w}_L$. The solid line shows the contribution of the moral principles to the $\alpha_{L,i}$ for each of the agents (in arbitrary order). The dotted line is the additional contribution 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).

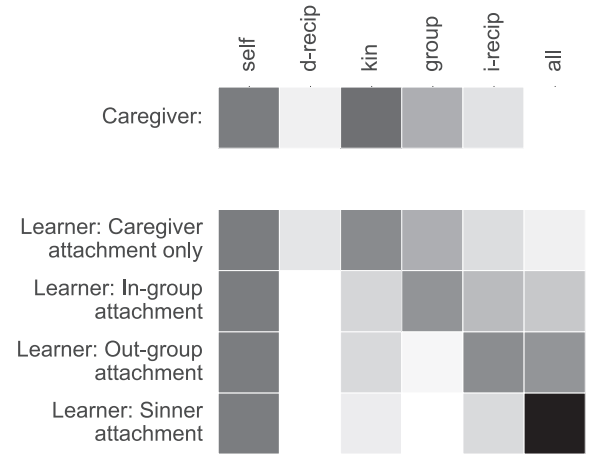

Fig. 6. Broadening a parochial moral theory through attachments and internal alignment. The caregiver and all other agents have parochial values (shown in the "Caregiver" row) which were inferred by the learner as in Fig. 3. When the learner only has a primitive attachment for the caregiver (like those shown in Fig. 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.

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

being victimized by crime), that experience may drive them toward a more parochial weighting of principles. Fig. 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 Fig. 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. 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 Fig. 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.

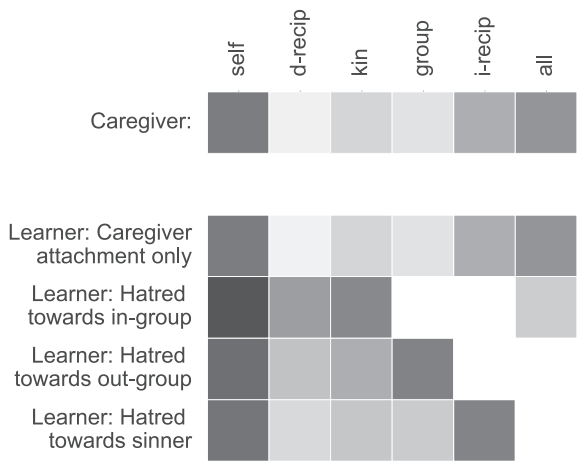

Fig. 7. Narrowing an impartial moral theory through feelings of hatred and internal alignment. The caregiver and all other agents have impartial values (shown in the "Caregiver" row) so change cannot occur through external alignment. These moral theories were inferred by the learner as in Fig. 3. When 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.

4.3. 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 behavior and judgment from the previous generation, infer the underlying moral theory (as in Fig. 3) and through value alignment, set the weights on their own moral theory (as in Fig. 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 (Fig. 8a), 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 (Fig. 8b).

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

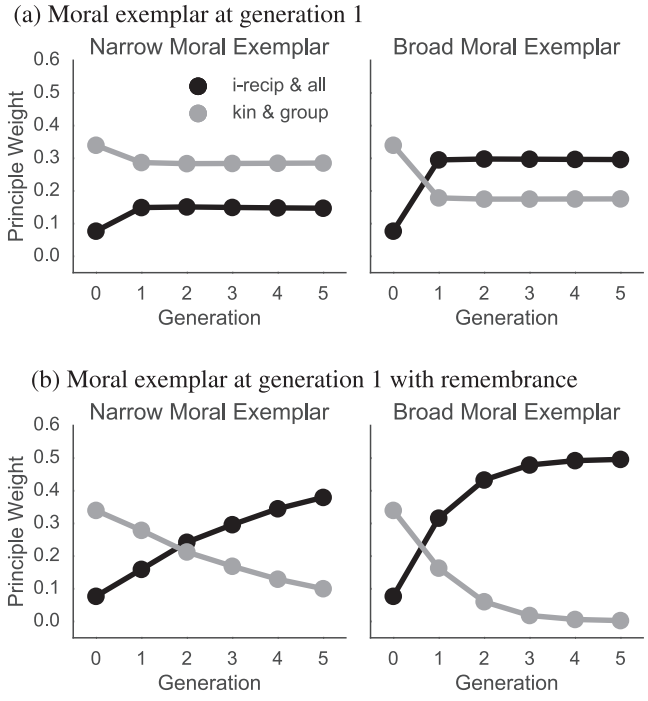

Fig. 8. Moral exemplars can rapidly reshape moral theories. When a moral exemplar with impartial values is introduced to parochial minded agents at generation 1 (a), the moral theories immediately adjust. There was a larger shift in moral theories when the moral exemplar was stronger (right) and affected 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.

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

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 Fig. 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. Fig. 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 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. Fig. 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.

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 (Fig. 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

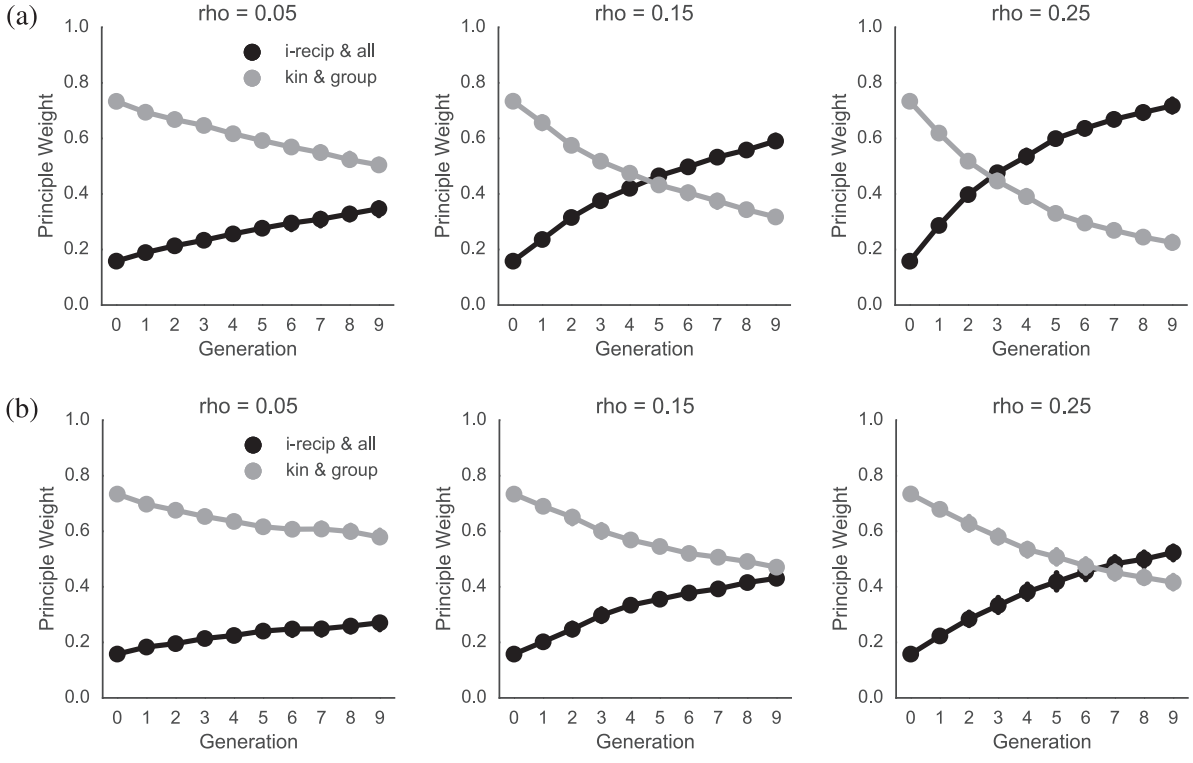

Fig. 9. Change in the average agent's weighting of parochial vs. impartial moral principles as a function of generation and the proportion of agents (ρ) that develop an attachment (ϕ) for another agent chosen (a) uniformly at random or (b) in proportion to their interaction frequency. The 0th generation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

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

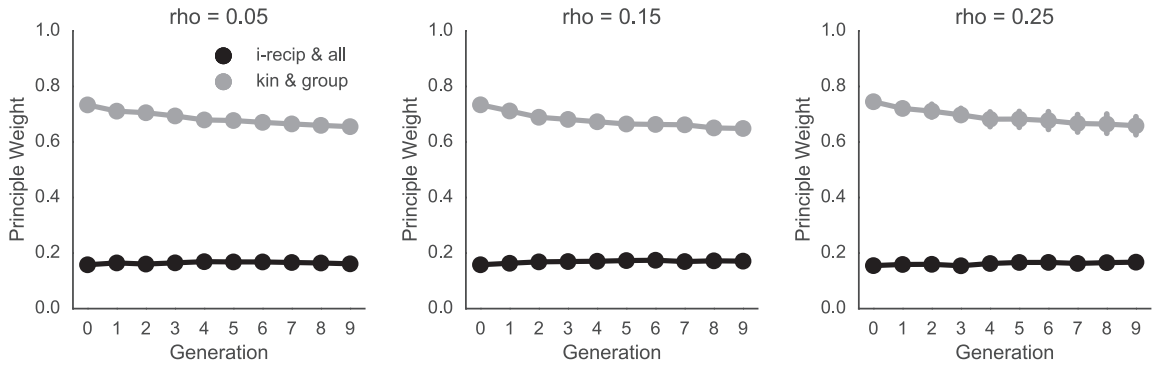

Fig. 10. Moral change from attachments critically depends on internal alignment. When simulations are run without internal alignment active during learning, there is no significant moral change towards impartial values no matter the proportion of agents (ρ) that develop an attachment for another agent.

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. Fig. 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 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.

5. 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

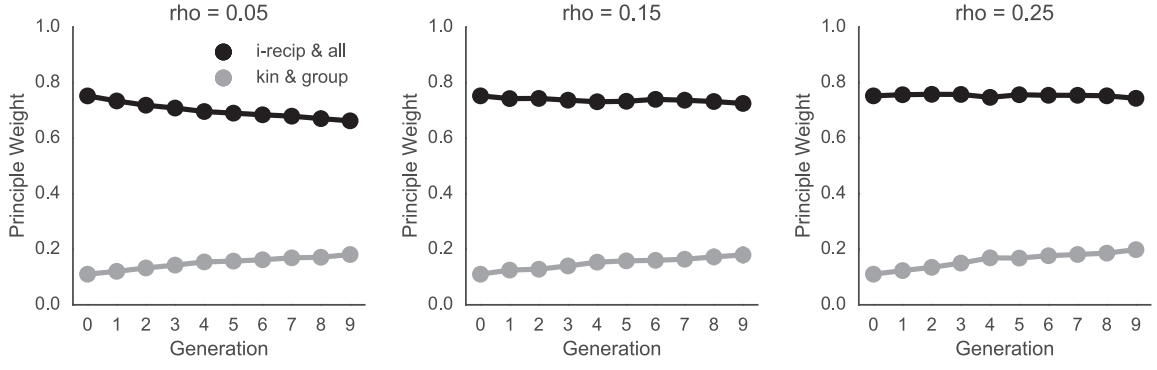

Fig. 11. Moral change towards impartial values is robust to in-group attachments. Agents started with an impartial moral theory but each generation developed attachments towards others with probability proportional to their interaction frequency. Thus most of these attachment were formed with kin and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

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

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 (Cushman, 2013; Nowak, 2006; Rand & Nowak, 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; 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 (Mikhail, 2011; Pinker, 2011; Pizarro, Detweiler-Bedell, & Bloom, 2006; Singer, 1981). 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 (Ayars & Nichols, 2017; Nichols et al., 2016) 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; Graham et al., 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

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

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 (Binmore, 1998; Rawls, 1971) or implemented in a more online fashion in individuals' "virtual bargaining" with each other (De Cote & Littman, 2008; Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016; Misyak, Melkonyan, Zeitoun, & Chater, 2014). 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 (Greene, 2014; Rand & Nowak, 2013). 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 (Bostrom, 2014; Wiener, 1960). 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 into 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.

Acknowledgements

We thank Fiery Cushman, Tobias Gerstenberg, Victor Kumar, Sydney Levine, Rachel Magid, Adam Morris, Peter Railton, Laura Schulz, and Tomer Ullman for discussions and ideas. MKW was supported by a Hertz Foundation Fellowship and NSF-GRFP. JBT was funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333. MKW, RS and JBT were supported by the Center for Brains, Minds and Machines (CBMM).

Appendix A. 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) 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_{norm}, and samples from the two reciprocity relationships $(P(f^{\text{d-recip}}, f^{\text{i-recip}}|H))$. In all simulations we used $\lambda = 1$ for $P(f^{\text{d-recip}})$, $p = 0.5$ for $P(f^{i-recip})$ and $\Sigma^g = \mathbf{I}$ for all g.

Appendix B. 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 (Kleiman-Weiner et al., 2015; Mikhail, 2007), these constraints could include restrictions on intention such as the doctrine of double effect or other specific forbidden actions (Haidt, 2007; Tetlock et al., 2000). 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 (Ayars & Nichols, 2017; Cushman, 2013; Nichols et al., 2016). 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 Eq. (2) for $U(s',a)$.

References

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

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

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

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.

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.

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

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., et al. (2015). The ontogeny of fairness in seven societies. *Nature*.

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

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

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.

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

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.

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.

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

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

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

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). Model s of morality. *Trends in Cognitive Sciences, 17*(8), 363–366.

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

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.

Festinger, L. (1962). *A theory of cog nitive 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.

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.

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. (n.d.). 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.

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* (Vol. 29, pp. 3909–3917). 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.

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.

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

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

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.
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., et al. (2013). Ontogeny of prosocial behavior across diverse societies. *Proceedings of the National Academy of Sciences, 110*(36), 14586–14591.
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.
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.
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*. .
Kohlberg, L. (1981). *The philosophy of moral development: Moral stages and the idea of justice*. Harper & Row.
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.
Lake, B., & Tenenbaum, J. (2010). Discovering structure by learning sparse graph. In *Proceedings of the 33rd annual cognitive science conference.*. .
Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature, 445*(7129), 727–731.
Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley.
Magid, R. W., & Schulz, L. E. (this issue). Moral alchemy: How love changes norms.
Malle, B. F., Moses, L. J., & Baldwin, D. A. (2001). *Intentions and intentionality: Foundations of social cognition*. MIT Press.
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*.
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.
Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H. Y. (2016). Rational learners and moral rules. *Mind & Language, 31*(5), 530–554.
Nagel, T. (1989). *The view from nowhere*. Oxford University Press.
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*. Cambridge, MA: MIT Press.
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. A. (2006). Five rules for the evolution of cooperation. *Science, 314*(5805), 1560–1563.
Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291–1298.
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.
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.
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.
Popper, K. S. (2012). *The open society and its enemies*. Routledge.
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.
Prince, A., & Smolensky, P. (2008). *Optimality theory: Constraint interaction in generative grammar*. John Wiley & Sons.
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., et al. (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.
Scanlon, T. M. (1975). Preference and urgency. *The Journal of Philosophy, 72*(19), 655–669.
Schäfer, M., Haun, D. B., & Tomasello, M. (2015). Fair is not fair everywhere. *Psychological Science* 0956797615586188.
Sen, A., & Hawthorn, G. (1988). *The standard of living*. Cambridge University Press.
Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General, 141*(2), 382.
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.
Singer, P. (1981). *The expanding circle*. Oxford: Clarendon Press.
Smetana, J. G. (2006). Social-cognitive domain theory: Consistencies and variations in childrens moral and social judgments. *Handbook of moral development* (pp. 119–153).
Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life, 9*(4), 371–386.
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.
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.
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).
Watanabe, T., Takezawa, M., Nakawake, Y., Kunimatsu, A., Yamasue, H., Nakamura, M., et al. (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.
Wiener, N. (1960). Some moral and technical consequences of automation. *Science, 131*(3410), 1355–1358.
Wright, J. C., & Bartsch, K. (2008). Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly*, 56–85 (1982-).
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.

Please cite this article in press as: Kleiman-Weiner, M., et al. Learning a commonsense moral theory. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.005

Constructing Social Preferences From Anticipated Judgments: When Impartial Inequity is Fair and Why?

Max Kleiman-Weiner[1] **(maxkw@mit.edu), Alex Shaw**[2] **(ashaw1@uchicago.edu)** & **Joshua B. Tenenbaum**[1] **(jbt@mit.edu)**
[1]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139
[2]Department of Psychology, University of Chicago, Chicago, IL 60637

Abstract

Successful and repeated cooperation requires fairly sharing the spoils of joint endeavors. Fair distribution is often done according to preferences for equitable outcomes even though strictly equitable outcomes can lead to inefficient waste. In addition to preferences about the outcome itself, decision makers are also sensitive to the attributions others might make about them as a result of their choice. We develop a novel mathematical model where decision makers turn their capacity to infer latent desires and beliefs from the behavior of others (theory-of-mind) towards themselves, anticipating the judgments others will make about them. Using this model we can construct a preference to be seen as impartial and integrate it with preferences for equitable and efficient outcomes. We test this model in two studies where the anticipated attribution of impartiality is ambiguous: when one agent is more deserving than the other and when unbiased procedures for distribution are made available. This model explains both participants' judgments about the partiality of others and their hypothetical decisions. Our model argues that people avoid inequity not only because they find it inherently undesirable, they also want to avoid being judged as partial.

Keywords: fairness, social cognition, theory-of-mind, decision making, Bayesian models

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 (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 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

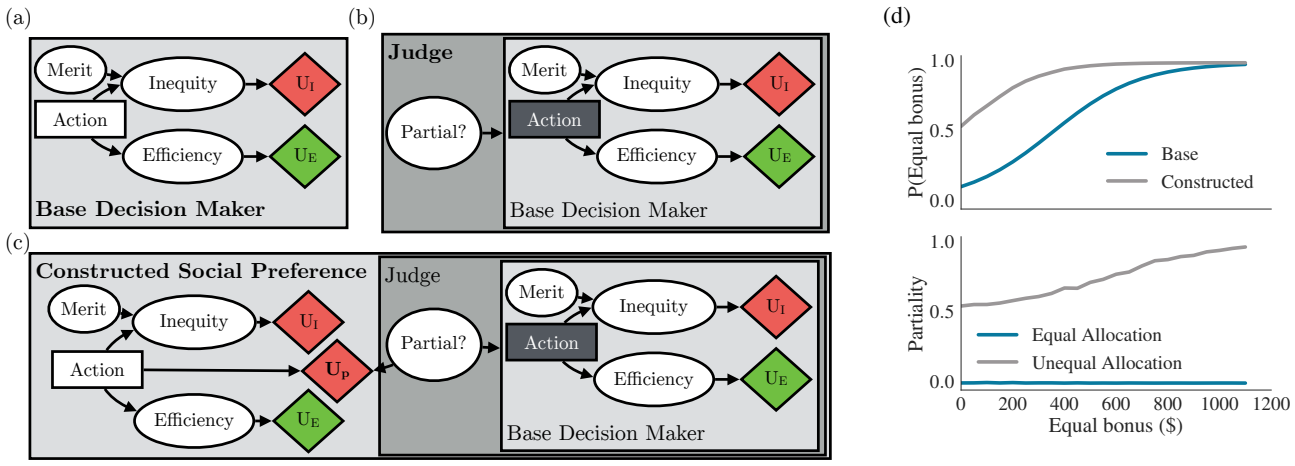

Figure 1: An influence diagram (ID) is a directed acyclic graph over three types of nodes: state nodes (circles), decision nodes (rectangles), and utility nodes (diamonds). Directed edges between nodes determine causal dependencies. State and utility nodes take values that depend on the values of their 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 (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 (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.

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.

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):

$$\frac{r_1}{\gamma_1} = \frac{r_2}{\gamma_2} = \ldots = \frac{r_N}{\gamma_N} \tag{1}$$

We transform these constraints into a measurement of inequity:

$$I(R, \boldsymbol{\gamma}) = \sum_{i \in N} \sum_{\substack{j \in N \\ j > i}} |\gamma_j r_i - \gamma_i r_j| \tag{2}$$

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:

$$\mathrm{EU}_{\texttt{base}}[a] = -\alpha_{IA} \mathrm{E}_a[I(R, \boldsymbol{\gamma})] + \sum_{i \in N} \alpha_i \mathrm{E}_a[r_i] \tag{3}$$

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:

$$P(a|\boldsymbol{\alpha}) \propto \exp(\beta * \mathrm{EU}[a]) \tag{4}$$

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, Gerstenberg, Levine, & Tenenbaum, 2015). The utility of the *Base Decision Maker* which is defined in equation (3) can be expressed graphically as the influence diagram shown in Figure 1a. The first term of equation (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:

$$P(\texttt{partial}, \boldsymbol{\alpha}|a) \propto P(a|\boldsymbol{\alpha}) P(\boldsymbol{\alpha}|\texttt{partial}) P(\texttt{partial}) \tag{5}$$

where $P(a|\boldsymbol{\alpha})$ is the model of action shown in equation (4) and the $\boldsymbol{\alpha}$ are then marginalized out to obtain a posterior on $P(\texttt{partial}|a)$. Figure 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*:

$$\mathrm{EU}_{\texttt{constructed}}[a] = \mathrm{EU}_{\texttt{base}}[a] - \alpha_{PA} P(\texttt{partial}|a) \tag{6}$$

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 1c show how the *Constructed Social Preference* is built on top of the *Judge* and *Base Decision 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 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

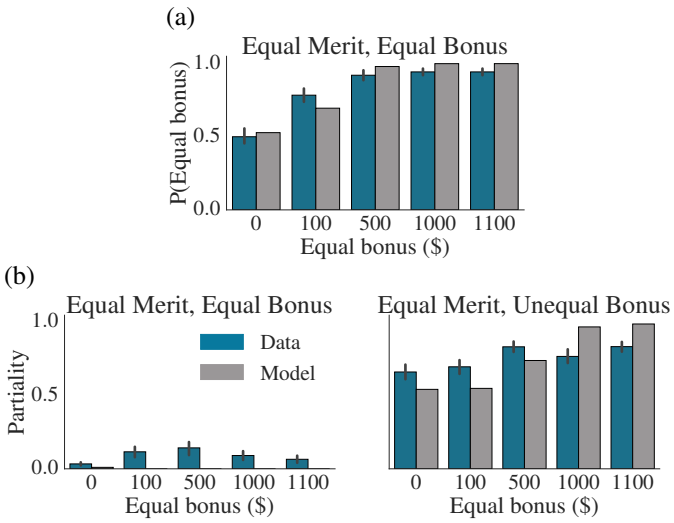

Figure 2: Empirical results and model predictions of (a) choices and (b) judgments of partiality for the trials in experiment 1 where both of the agents were equally meritorious. Trials with no gray bar indicate the model predicted near 0. Error bars are the standard error of the mean.

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.

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

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 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 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 dis-

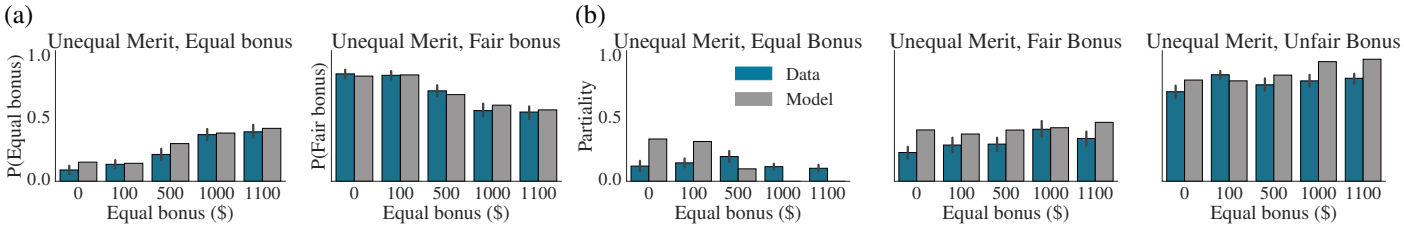

Figure 3: Empirical results and model predictions of (a) choices and (b) judgments of partiality for the trials in experiment 1 where one of the agents was more meritorious than the other. Trials with no gray bar indicate the model predicted near 0. A "fair bonus" was when the decision maker gave the 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.

tributions. 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.

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 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 5 shows the quantitative correlation 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 ($R^2 = 0.96$).

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

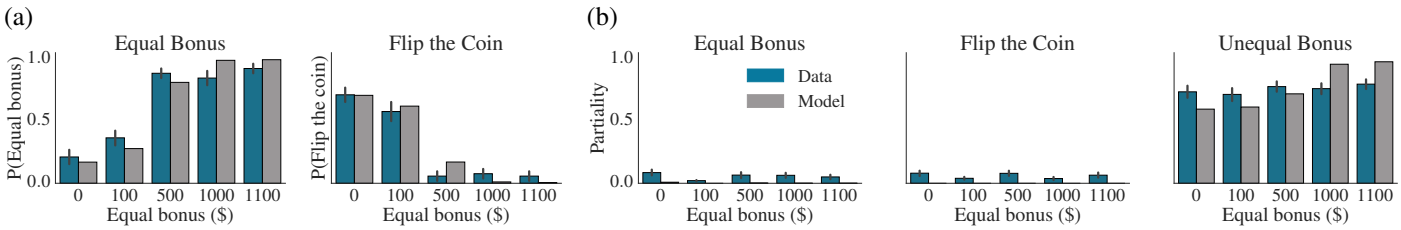

Figure 4: Empirical results and model predictions of (a) choices and (b) judgments of partiality for experiment 2 which introduced the option to flip a fair coin to decide the allocation of the unequal bonuses. Trials with no gray bar indicate the model predicted near 0. Error bars are the standard error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

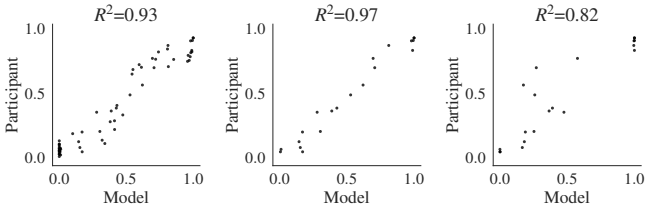

Figure 5: Quantification of model performance. Each point represents the model prediction and participant judgment for a single condition. For better fitting models the points will lie close to the $y = x$ diagonal. (left) The full model compared including both decision and judgment data. (middle) The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

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.

Acknowledgement This work was supported by a Hertz Foundation Fellowship, NSF-GRFP, the Center for Brains, Minds and Machines (CBMM), NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333.

References

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

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

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

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

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

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.

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.

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.

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

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

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

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.*

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

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Shaw, A. (2013). Beyond "to share or not to share" 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.

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.

Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.