UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

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

Permalink
https://escholarship.org/uc/item/74m9s697

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 39(0)

Authors
Kleiman-Weiner, Max
Shaw, Alex
Tenenbaum, Joshua B.

Publication Date
2017

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

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

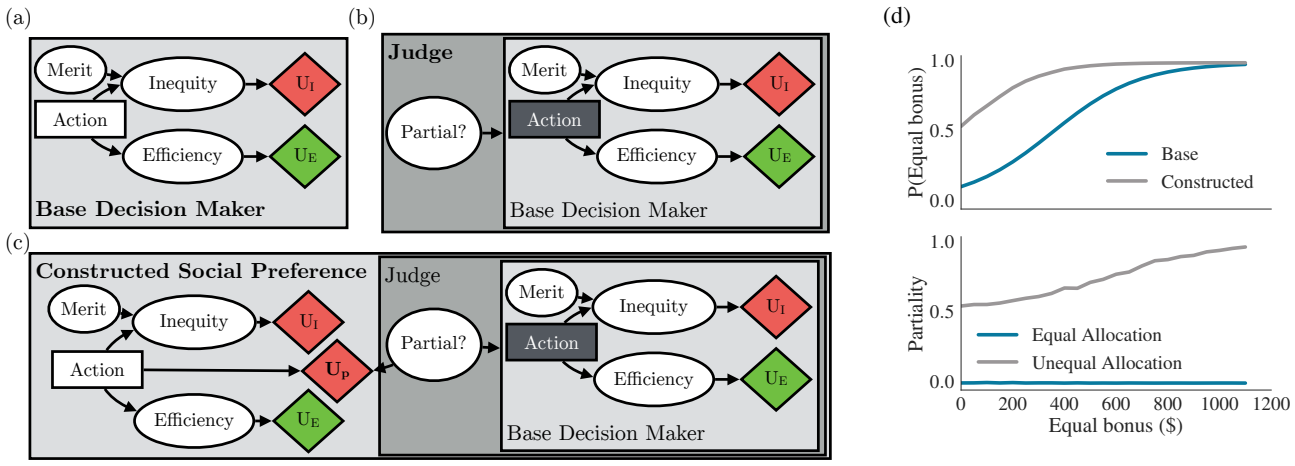

Figure 1: An influence diagram (ID) is a directed acyclic graph over three types of nodes: state nodes (circles), decision nodes (rectangles), and utility nodes (diamonds). Directed edges between nodes determine causal dependencies. State and utility nodes take values that depend on the values of their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

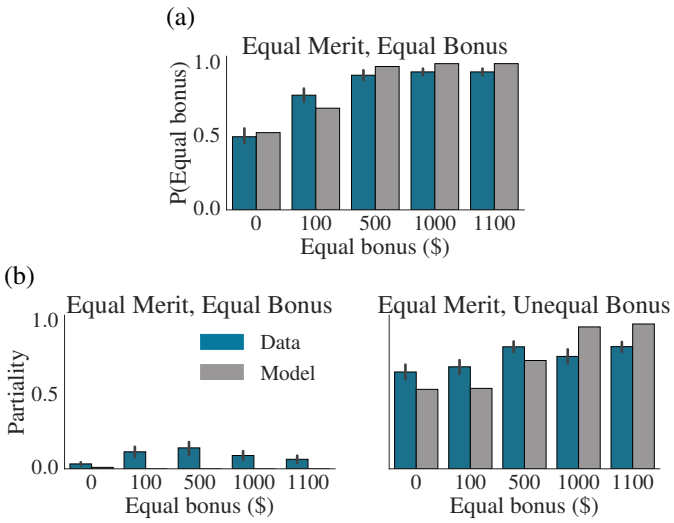

Figure 2: Empirical results and model predictions of (a) choices and (b) judgments of partiality for the trials in experiment 1 where both of the agents were equally meritorious. Trials with no gray bar indicate the model predicted near 0. Error bars are the standard error of the mean.

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-

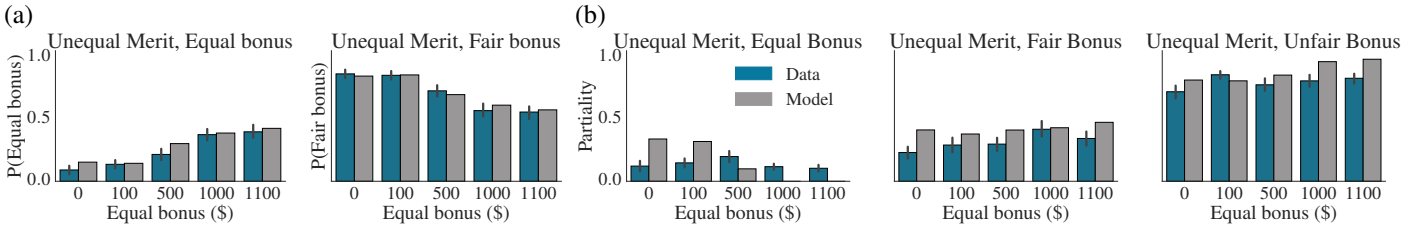

Figure 3: Empirical results and model predictions of (a) choices and (b) judgments of partiality for the trials in experiment 1 where one of the agents was more meritorious than the other. Trials with no gray bar indicate the model predicted near 0. A "fair bonus" was when the decision maker gave the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

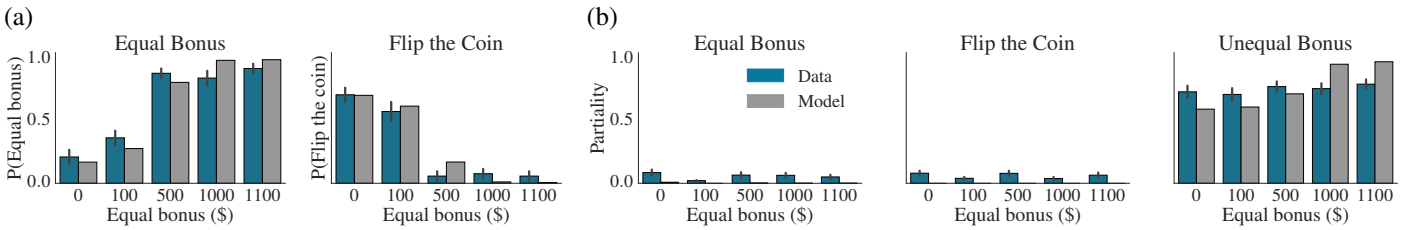

Figure 4: Empirical results and model predictions of (a) choices and (b) judgments of partiality for experiment 2 which introduced the option to flip a fair coin to decide the allocation of the unequal bonuses. Trials with no gray bar indicate the model predicted near 0. Error bars are the standard error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

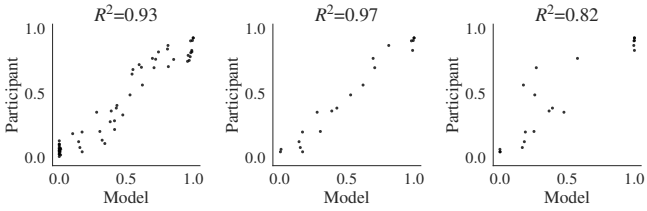

Figure 5: Quantification of model performance. Each point represents the model prediction and participant judgment for a single condition. For better fitting models the points will lie close to the $y = x$ diagonal. (left) The full model compared including both decision and judgment data. (middle) The 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.