Reasoning about social preferences with uncertain beliefs

Isaac Davis (isaac.davis@yale.edu), Ryan Carlson (ryan.carlson@yale.edu)
Yarrow Dunham (yarrow.dunham@yale.edu), Julian Jara-Ettinger (julian.jara-ettinger@yale.edu)
Department of Psychology, Yale University. 2 Hillhouse Avenue, New Haven, CT 06520 USA

Abstract

We propose a computational model of social preference judgments that accounts for the degree of an agents' uncertainty about the preferences of others. Underlying this model is the principle that, in the face of social uncertainty, people interpret social agents' behavior under an assumption of expected utility maximization. We evaluate our model in two experiments which each test a different kind of social preference reasoning: predicting social choices given information about social preferences, and inferring social preferences after observing social choices. The results support our model and highlight how uncertainty influences our social judgments.

Keywords: Social Inference; Theory of Mind; Computational Modeling

Introduction

Imagine Pat arrives in the break room at work to find a box of donuts with three donuts left: one jam donut, and two glazed donuts. Pat's co-worker, Sam, doesn't have a break for another few minutes, so Pat gets first pick. Which donut will Pat pick? The answer likely depends in part on Pat's preferences for donuts, but also on Pat's knowledge of Sam's preferences, and how much Pat cares about Sam's preferences relative to his own. For example, if Pat strongly prefers jam donuts and believes that Sam does too, but he doesn't particularly care about Sam, then Pat might take the jam donut without further thought. On the other hand, if Pat really cares about Sam, he might be more inclined to take the less preferred donut and leave the jam one for Sam. In this context, we can interpret the degree to which Pat cares about Sam (positively or negatively) as Pat's *social preference* toward Sam.

In social life, people exhibit a wide range of social preferences (Fehr & Fischbacher, 2002; McClintock & Allison, 1989)—such that some people tend to forgo their own self-interest to benefit others, and some people tend to prioritize their own interests over the interests of others. As such, making accurate inferences about peoples' social preferences can be crucial for navigating a social environment. How do people accomplish this feat? Previous research suggests that people reason about social decisions and social preferences by drawing on Theory of Mind (Ullman et al., 2009; Frith & Frith, 2012), the capacity to interpret and predict other people's behavior by relying on representations of their mental states. In particular, when predicting social decisions, people are sensitive to the decider's preferences over outcomes, beliefs about others' preferences over outcomes, and social preferences toward those affected by the decision (Van Doesum et al., 2013). Conversely, people are similarly sensitive to preferences and beliefs when reasoning in the opposite direction, i.e. observing a decision and inferring the status of the decider's social preference toward those affected by the decision (whom we shall henceforth refer to as the "receivers"); (Carlson & Zaki, 2018; Jara-Ettinger et al., 2015; Jern & Kemp, 2014). In previous work, Jern & Kemp (2014) leveraged a utility-based model to predict people's inferences in situations like these.

In many cases, however, the decider may be uncertain about the receiver's preferences over outcomes. In the above example, suppose that Pat cares a lot about Sam, but does not know which kind of donut Sam prefers. In such cases, research suggests that people strongly associate "prosocial" preferences with actions that leave the receiver a choice—a phenomenon known as "social mindfulness" (Van Doesum et al., 2013). In this example, we might expect Pat to choose a glazed donut, thus allowing Sam to pick whichever donut he prefers more. Conversely, if we observe that Pat does in fact pick one of the glazed donuts, we might infer that Pat cares positively about Sam. Thus, in situations where preferences are uncertain, prosociality can be best achieved through social mindfulness – that is, gifting others the power to choose their own outcome.

In this paper, we propose that both social mindfulness and social preferences can be explained through a single utility based model by accounting for the decider's degree of uncertainty about the receiver's preferences. This model builds on prior work (Jern & Kemp, 2014), and advances this work by integrating the role of uncertainty into social reasoning. Specifically, here we model the fact that deciders often have varying degrees of uncertainty about a receiver's utility. By doing so, our model captures a broader set of social predictions and inferences over varying degrees of preference uncertainty, which we validate with a pair of novel experiments. In addition, we propose that social mindfulness—the well-documented phenomenon in social psychology in which people take actions that preserve others' ability to choose between different options (Van Doesum et al., 2013)—can be understood as a special case of a more general kind of utility-based social inference.

Computational Framework

Our computational framework takes as a starting point the idea that social inference is structured around an assumption that agents act rationally to maximize utilities—the difference between the costs they incur and the rewards they obtain (Gergely & Csibra, 2003; Jara-Ettinger et al., 2016; Jern et al., 2017; Lucas et al., 2014). Through this assumption, observers can break down people's actions into judgments about their underlying preferences and abilities via Bayesian inference (or some approximation thereof), enabling them to make a range of social inferences including determining other people's goals (Baker et al., 2009), beliefs (Baker et al., 2017), competence (Jara-Ettinger et al., 2020), and social intentions (Ullman et al., 2009).

Generative model

We focus on a general form of a decision problem that is common in the social mindfulness and social relationships literature (e.g., Van Doesum et al. 2013): a "giver" agent G starts with an initial allotment of "treats" of 2 different types, and must give a fixed total number of those treats to a "receiver" agent R. Following previous work (Jern & Kemp, 2014), we model G's overall utility as a weighted sum of two utility functions $U = w_G U_G + w_R U_R$. This captures the notion that G may care positively or negatively about R's utility. The term U_G denotes G's *direct* utility—this is the value G personally derives from the outcome of G's choice.

The term U_R denotes G's *belief* about the utility that R derives from the outcome of G's choice. The pair of weights (w_G, w_R) capture G's *social preference* towards R (Murphy & Ackermann, 2011). We restrict w_G and w_R to values in the interval $[-1,1]$. For example, $(w_G, w_R) = (1,0)$ captures a purely self-interested agent whose overall utility is solely dependent on their own direct utility, while $(w_G, w_R) = (0,1)$ captures a purely altruistic G whose overall utility is solely dependent on R's perceived direct utility. Negative values of w_R correspond to a negative social preference towards R, i.e. G's total utility *decreases* as R's increases. Let D denote G's decision problem, the outcome of which will directly affect both G and R. Suppose that G knows R's utility function U_R. In this case, the total utility G derives from decision $d \in D$ is

$$U(d) = w_G U_G(d) + w_R U_R(d) \tag{1}$$

In many cases, however, G may be uncertain about R's utility. Indeed, G may be completely uncertain (i.e. G knows nothing about what R prefers) or partially uncertain (e.g. G knows R prefers A over B, but does not know the strength of that preference). In order to capture this uncertainty, we represent G's beliefs as a probability distribution $P_R(U)$ over possible direct utility functions, rather than a single utility function U_R. In this case, the total utility G derives from decision d depends on R's *expected* utility, based on G's uncertain beliefs. Thus, our model defines G's total utility as

$$U(d) = w_G U_G(d) + w_R E[U_R(d)] \tag{2}$$

where the expectation is taken with respect to G's belief distribution $P_R(U)$. This belief distribution reflects the level of G's knowledge or uncertainty about R's preferences: if G is completely certain, the distribution is a point mass on one particular U_R. If G is completely uncertain then $P_R(U)$ is a (not-necessarily uniform) prior distribution. If G knows a non-zero amount (e.g. that R prefers A over B by an unknown degree) then $P_R(U)$ will reflect a prior distribution conditioned on this knowledge (e.g. a distribution over all U_R's for which $U_R(A) > U_R(B)$).

Given the total utility function defined in equation (2), we make the standard assumption that G chooses d probabilistically in proportion with its utility $U(d)$, using a softmax function (concentration parameter $\beta = 3$) to convert utilities into decision probabilities.

Direct utility functions

In these studies, G must decide how to allocate quantities of two types of treats (brownies and cupcakes) between G and R. The quantitative nature of this task, and the fact that the resources being allocated likely have a high rate of satiation (i.e. one's preference for cupcakes may drop significantly immediately after eating one or more cupcakes) suggest that the direct utility functions are likely to show discounting Baucells & Sarin (2010). For theoretical reasons (explained further in the results section), this discounting effect is important for capturing human inferences in tasks like these, and we validate this by comparing our predictions against those of an alternate model with no discounting.

We can capture this effect with a discount parameter $0 < \rho < 1$: if G receives utility U from eating a brownie, G will derive $\rho * U$ utility from a second brownie, $\rho^2 * U$ utility from a third brownie, and so on. For each agent, we assume a baseline utility value for each treat type (u_b, u_c), and define the total direct utility of allocation [B brownies, C cupcakes] as

$$U^{direct}([B,C]) = u_b * D(B,\rho) + u_c * D(C,\rho) \tag{3}$$

where $D(X,\rho)$ applies the discount function described above to quantity X with discount rate ρ. While the discount value ρ introduces an additional parameter to the model, we opted to integrate this parameter out of the model under a $Beta(2,1)$ prior distribution, rather than attempting to estimate the value of this parameter from human data. Intuitively, this is equivalent to modeling inference under the assumption that agents' direct utility functions are discounted, but the true discount rate is unknown. To demonstrate the importance of discounting, we compare our results against an alternate model which differs only by removing the discounting effect (i.e. fixing $\rho = 1$).

Inference

The generative model above allows us to capture various kinds of inferences about social choices and social preferences under varying degrees of uncertainty. We focus on two kinds of inferences. First, given some information P_G about

G's direct preferences, G's beliefs P_R about R, and some information P_W about G's social preference towards R, to predict G's decision $d \in D$, where D is some decision problem that affects both G and R. The generative model defined above provides the decision probability $P(d|U_G, P_R, (w_G, w_R))$. By placing prior distributions over U_G, w_G, and w_R, a Bayesian observer can integrate these parameters over an appropriate range (determined by P_G and P_W) to compute the probability of each possible decision. We initially used a uniform prior over all model parameters. However, initial pilot studies suggested that participants were less inclined overall to attribute an antisocial social preference to G (i.e. one for which $w_R < 0$, so that G's utility *increases* as R's *decreases*), and were generally biased to report that G attributed a positive or at least non-negative weight to R's direct utility. For this reason, our final predictions use an asymmetric Beta prior (re-scaled to the range $(-1,1)$) over the parameter w_R. The degree to which this distribution is skewed towards positive values is determined by a "niceness bias" parameter ν, which we fitted to pilot data prior to analysis.[1]

In the other direction, if we do not know anything about G's social preference, observing G's decision d may help reveal that information. In particular, we can infer a posterior distribution over G's social preference according to Bayes' rule:

$$P((w_G, w_R)|P_G, P_R, d) \propto P(d|(w_G, w_R), P_G, P_R)P((w_G, w_R)) \quad (4)$$

The likelihood term $P(d|(w_G, w_R), P_G, P_R)$ is determined by the generative model + marginalization as above, while $P((w_G, w_R))$ is the prior distribution over social preferences. To compute direct utilities, we use the discounting function described in the previous section, and integrate out the discount parameter ρ over a $Beta(2,1)$ prior. As we show in our experimental task, the discounting factor can affect our model's decisions and inferences in an important way. We therefore created a second baseline model that was identical to our main model with the difference that it had no discounting.

Experiments

We conducted two studies, one for each form of inference described in the previous section.

Study 1

Participants 40 adult participants with US-based IP addresses were recruited via Amazon Mechanical Turk (mean age=36.3, S.D.=10.4). 6 Additional participants were recruited but excluded for failing one or more of 6 comprehension check questions.

Stimuli Stimuli consisted of 27 trials divided across 9 distinct scenarios. Each scenario depicts a named giver (G) and receiver agent (R), G's initial endowment of two kinds of treats (brownies and cupcakes), and a fixed number of treats that G must give to R (we fixed this number to 2 for all trials). Importantly, the scenario is explained so that R will get to keep all of the treats they receive from G. This differs from prior work on social mindfulness Van Doesum et al. (2013), where the receiver chooses only one of the objects they receive from G. Each scenario also depicts information about G's preferences and beliefs. G's preferences are represented using a thought bubble containing a statement of the form "I prefer [brownies/cupcakes]" or "I like brownies and cupcakes equally." G's beliefs are represented using a different thought bubble containing a statement of the form "R prefers [brownies/cupcakes]" or "I don't know what R prefers." Examples of these stimuli are shown in Figure 1a.

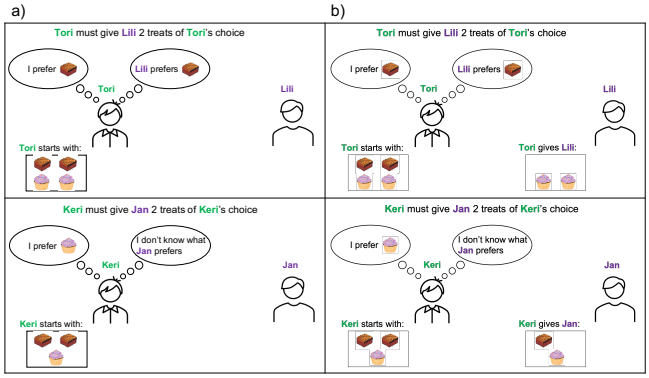

Figure 1: Panel a) depicts two examples of stimuli from Study 1. Panel b) depicts two examples of stimuli from Study 2. Note that the receiver agent gets to keep all of the treats given by the giver agent

The 9 scenarios were divided into three "blocks." Within each block, we fixed G's initial endowment and personal preference, but varied G's belief about R's preference between the three possibilities described above. This allows us to assess how G's knowledge or uncertainty interacted with G's social preference to influence participants' predictions. For each scenario, participants are asked to predict G's decision under three different hypothetical conditions:

1. (Selfish) "Suppose G places higher value on his/her own preferences than R's."

2. (Altruistic) "Suppose G places higher value on R's preferences than his/her own."

3. (Egalitarian) "Suppose G places about the same value on R's preferences and his/her own."

Note that this does not specify G's exact social preference towards R, only a general range for that social preference. These three conditions, in conjunction with the 9 different scenarios, result in a total of 27 trials for Study 1.

[1] The exact form of this prior is $w_R \sim 2 * Beta(\nu, 1) - 1$. We used $\nu = 2.5$, as estimated from earlier pilot data, to generate predictions for the final study.

Procedure Participants were first shown a series of instructions explaining the general context and how to interpret the information in each stimulus picture. After the initial instructions, participants were given two chances to pass a 6-question comprehension check ensuring that they were able to correctly interpret the stimuli. Participants who failed one or more questions on both tries were excluded from the study.

Upon passing the comprehension check, participants were then shown all 27 trials. The 9 different scenarios were presented in a random order, and within each scenario, the three "social preference" conditions were also presented in a random order. For each condition, participants selected the treat allocation they believed G was most likely to give. For the [2B, 1C] initial endowment scenarios, the options were [1B, 1C] or [2B, 0C]; for the [2B, 2C] initial endowment scenarios, the options were [1B, 1C], [2B, 0C], or [0B, 2C]. These options were shown as pictures with a verbal description below each picture.

Results As preregistered [2], for each scenario and condition we computed the proportion of participants who chose each possible action, and compared these numbers against the predicted action probabilities generated by the computational model. Figure 2 depicts results of this study by social preference condition.

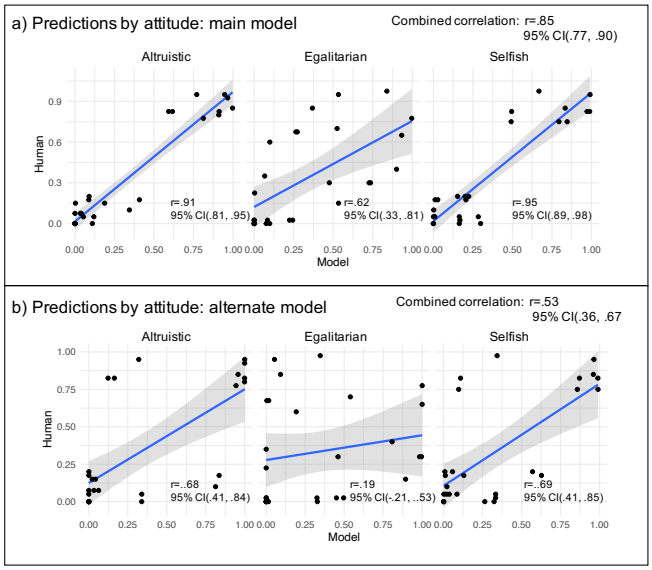

Figure 2: Comparison of model predictions against human data, separated by "social preference" condition. Each point represents one action probability in one trial, with model predictions on the x axis and aggregate human judgments on the y axis. Grey bands indicate 95% confidence intervals in a linear regression. Panel a) depicts results from the main model, panel b) depicts results from the alternate (no-discounting) model

In aggregate, the model predictions were correlated with participant judgments, $r = .85$ (95% CI: .77, .90). While accuracy was quite high in two of the three conditions, the degree of mismatch between model and data was higher in the "egalitarian" condition ($r = .62$, CI (.33, .81)): correlations for the "selfish" and "altruistic" conditions were significantly higher (selfish: $r = .91$, CI (.81, .95); altruistic: $r = .95$, CI (.89, .98)). Furthermore, participants' responses in the "egalitarian" condition were frequently identical with their responses in the "altruistic" condition. One possibility for this result is that participants found it easy to identify the best choice in the "selfish" and "altruistic" conditions, but struggled more in determining what choice would be equally utility-maximizing for both agents, defaulting to a more altruistic choice. This decision is particularly difficult given that the egalitarian condition may be inherently more ambiguous, as it entails a wider range of possible configurations for model parameters (e.g. G may care about R and herself equally, but very strongly prefer one treat over the other, which leads G to behave as though G is selfish). In future work we will seek to further disentangle and clarify how we reason about egalitarian social preferences.

In contrast to the strong fit between participants and our model, the baseline model with no discounting showed a substantially lower correlation with participants (combined: r=.53, 95% CI(.36, .67), demonstrating the importance of diminishing marginal returns in the utility function. Correlation was significantly lower with the alternate model both combined and within social preference conditions.

Figure 3 shows a pair conditions that reveal people's sensitivity to G's level of knowledge, particularly in the altruistic condition, where participants predict that G will give one item of each if they do not know R's preference. Critically, this effect only appears when utilities are discounted. Otherwise, the average expected reward for any combination of items is the same (given that R is equally likely to like any of them).

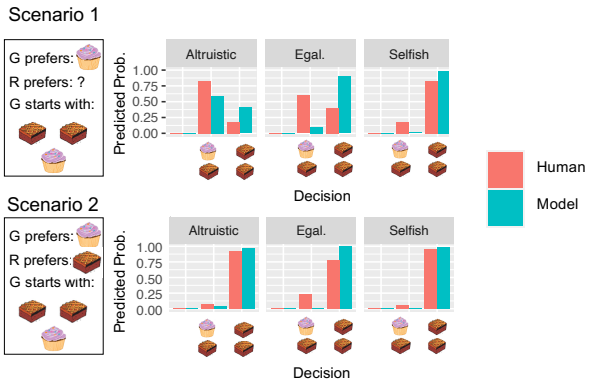

Figure 3: Comparison of predictions between two scenarios from Study 1, which differ only in G's knowledge state.

Study 2

Participants 40 adult participants with US-based IP addresses were recruited via Amazon Mechanical Turk (mean

[2]Pregistrations: osf.io/zdb3y (study 1), osf.io/23rq7 (study 2)

age=40.1, S.D.=11.8). 4 Additional participants were recruited but excluded for failing one or more of 6 comprehension check questions.

Stimuli Stimuli were almost identical to Study 1 stimuli, with the key difference being that each stimulus showed G's decision. Examples of these stimuli are shown in Figure 1b.

We started with the same 9 scenarios from Study 1, and for each scenario, created a separate trial for each possible action. The first three scenarios (with initial endowment of [2B, 1C]) had only 2 possible actions, while the remaining scenarios each had three, yielding $3 \times 2 + 6 * 3 = 24$ total trials.

Procedure Participants were first shown a series of instructions explaining how to interpret the stimulus picture. Participants were then given two chances to pass a 6-question comprehension check; participants who failed at least one question on both tries were excluded from the study. Upon passing the comprehension check, each participant was shown all 24 trials in a random order. For each trial, participants were asked two questions:

1. "On a scale from 0 to 5, how much do you think G values his/her own preferences?"[3]

2. "On a scale from -5 to 5, how much do you think G values or disvalues R's preferences?" This question was accompanied by a note clarifying that a negative value implies that G *dislikes* when R gets what they prefer.

Participants provided their answers by adjusting a sliding scale on the screen. Note that this experiment is similar to the social relationship studies performed in Jern & Kemp (2014), but is importantly different in two ways. First, we include an "uncertainty" condition in which G does not know what R prefers. Second, whereas Jern & Kemp ask participants to provide a discrete label describing the two agents' relationship (friend, enemy, or stranger), we ask participants to provide numerical estimates about each weight. This enables us to capture more graded inferences, e.g. participants can infer that G cares a lot about R in one trial, and infer that G cares about R only somewhat (but still positively) in a different trial.

Results Participants provided separate judgments about self-weight (w_G in the model) and other-weight (w_R in the model) for each of 24 trials, yielding 48 total inferences. As preregistered, we Z-scored responses within participant (separately for each of the two parameters) and then averaged them. We then compared these averages against the mean output of the model for each inference (Z-scored across trials). Model predictions were highly correlated with participant responses, $r = .89$, 95% CI(.81, .94). Figure 4 shows scatterplot comparisons between model predictions and participant judgments for both the main and alternate model. The alternate model (with no discounting) performed substantially worse than the main model (combined correlation r=.73, 95% CI(.57, .84)).

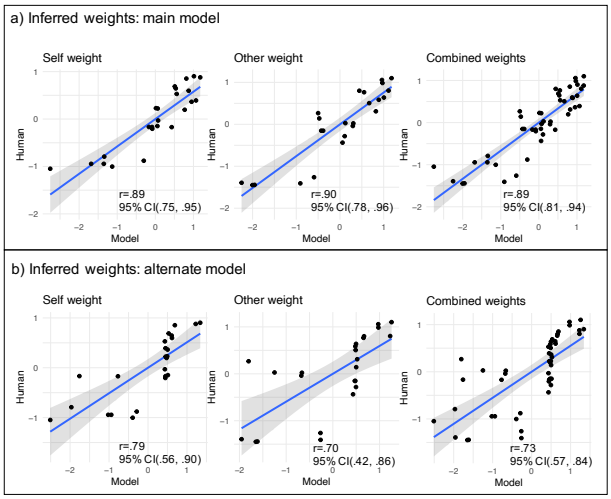

Figure 4: Comparison of model predictions against human data. Each point represents the average weight judgment for a single trial, with model predictions on the x axis and aggregate human judgments on the y axis. Grey bands indicate 95% confidence intervals in a linear regression. Panel a) depicts results from the main model, panel b) depicts results from the alternative (no-discounting) model

As in Study 1, Study 2 also reveals how G's uncertainty affects participant judgments. For example, in trials 1.1 and 1.2, where G does not know R's preference, both participants and the model infer that G is altruistic when G gives [1B, 1C], and selfish when G gives [2B, 0C] (see Figure 5). In trials 1.3 and 1.4, however, which are identical to 1.1 and 1.2 except that G knows R prefers B>C, participants and the model now infer that G is weakly altruistic or egalitarian when G gives [2B, 0C]. Thus, as in Study 1, this reveals how participants are sensitive to G's knowledge or uncertainty when making judgments about social preferences.

Discussion

In our everyday social experience, we frequently make inferences about the social preferences and choices of those around us. A growing body of previous work shows that these everyday social inferences reflect a sensitivity to agents' feelings and preferences, unified by an underlying assumption of utility maximization (Jara-Ettinger et al., 2016; Jern et al., 2017). We build on this work by proposing that agents' social inferences are sensitive to the decision-maker's degree of certainty or uncertainty about the preferences of others in a way that reflects an underlying utility calculus.

Across two unique experiments, we found converging support for this proposal. When a decider's social preference

[3]We initially allowed the first question to range from -5 to 5. However, we were concerned that participants would be confused by the notion of placing a negative weight on one's own direct utility. Since we did not consider any trials in which a negative self-weight would be relevant, we changed the first question to only range from 0 to 5

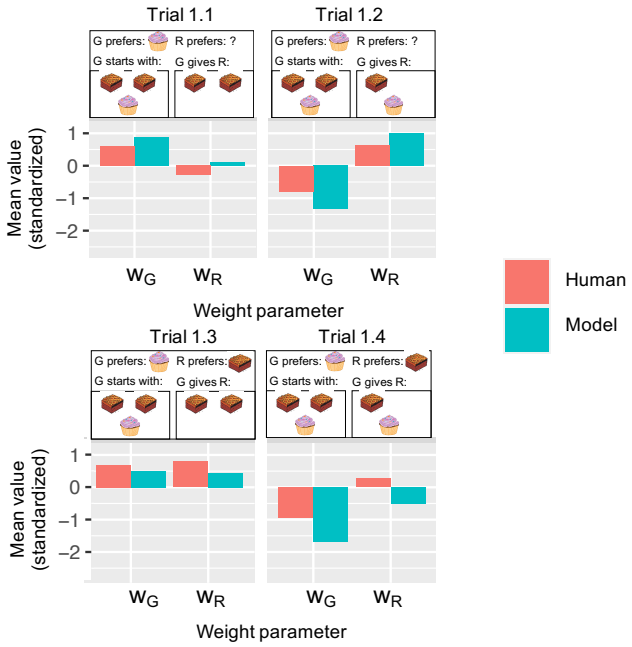

Figure 5: Comparison of inferred social preferences between the first and second pair of conditions in Study 2, which differ only in G's knowledge state.

was known (e.g., that they are self-interested; Study 1), perceivers' predictions about the decider's most likely action strongly tracked with a utility based model. Moreover, we find support for our account when perceivers confront the inverse problem as well. When a decider's action was known (Study 2), perceivers' inferences about the decider's social preference again strongly tracked with our model. Taken together, these findings offer convergent evidence for utility based models of social inference, and highlight the promise of incorporating preference uncertainty into such models. Furthermore, these results suggest that a utility-based model which incorporates graded uncertainty can account for patterns in human social judgments that are commonly reported in literature on both social relationships (Jern & Kemp, 2014) and social mindfulness (Van Doesum et al., 2013).

While these results are promising, they also contain important nuances. While our model makes fine-grained predictions about how people reason about self-interested and altruistic preferences, it does less well at capturing perceivers' predictions about what action a decider will take when they hold an egalitarian preference–that is, when they care equally about their own and another person's preferences. Our results from Study 1 suggest that, in these cases, perceivers interpret egalitarian preferences in a way that is difficult to distinguish from an altruistic preference. Curiously, participants in Study 2 had no issue *inferring* an egalitarian preference (i.e. $w_G \approx w_R$) in appropriate trials. It is possible that people's reasoning in egalitarian contexts works in a different way than our model captures. For example, people may generally consider egalitarianism in an iterative context (e.g. "I take one now, then you take one next time"). If this were the case, predicting the decision for a single trial would not reveal the participant's reasoning. Further research is required to disentangle these possibilities and test whether the results could have emerged from task artifacts. Additionally, many of the egalitarian trial conditions contain insufficient information to make a conclusive inference. For example, if G and R both prefer brownies, and there is only one brownie to allocate, then there is no clear "egalitarian" allocation, only a selfish allocation (G keeps the brownie) and an altruistic allocation (G gives R the brownie). In a case like this, the model randomizes, and predicts either action with roughly equal probability, but it is possible that participants draw on other expectations or information for under-determined cases like these.

A further limitation of the current work is that it considers individual instances of social decisions outside of any broader social context. In reality, there are numerous other factors that could shape our social inferences as well. One important factor to consider in future work is reputation: in particular, our social decisions are often observed by others in our environment. In such cases, we may care not only about the direct utilities of those affected by our decisions, but also the opinions of those who observed (but may not be directly impacted by) those decisions. Existing work suggests that social behavior is highly influenced by reputational concerns (Ariely et al., 2009), and perceivers are sensitive to such influences when making inferences about why people engaged in prosocial behavior (Barasch et al., 2014; Carlson & Zaki, 2018). It is therefore important to extend our model to account for reputational factors or other kinds of social pressures.

Beyond situational factors, another crucial consideration lies within perceivers themselves. People form rich models of their social worlds that contain expectations for the social preferences and actions of others. In our current model, we represented part of these expectations with a non-uniform prior over w_R (the weight that G assigns to R's utility). This prior encoded a "niceness bias," reflecting an underlying expectation that social agents are generally more likely to weight others' utilities positively rather than negatively. However, group identities and intergroup dynamics can significantly influence our perceptions and expectations of how social agents treat each other (Rhodes, 2013). In the context of our model, this suggests that people's prior expectations over utility weights may be highly dependent on social information like group identity. For example, if we know that G and R belong to the same in-group, we might have stronger expectations that G will make an altruistic or egalitarian decision than if R belongs to an out-group. Conversely, if we do not know G and R's group membership status, observing how G treats R may provide information about that status. Accounting for this group structure, and other relevant contextual factors, are important future extensions for our framework.

To conclude, we proposed an extension to existing utility-

based models of social inference by considering how the decider's certainty or uncertainty influences perceivers' judgments about social preferences and social choices. We demonstrate that perceiver's judgments are sensitive to this uncertainty in a way that tracks the expected utility calculations of our model, for both predicting social choices and inferring social preferences. This constitutes an important step toward a broader understanding of our everyday social reasoning.

Acknowledgments

We thank the members of the Yale Computational Social Cognition Lab for feedback and support. This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–55.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

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

Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? on the signal value of emotion in altruistic behavior. *Journal of personality and social psychology*, *107*(3), 393.

Baucells, M., & Sarin, R. K. (2010). Predicting utility under satiation and habit formation. *Management Science*, *56*(2), 286–301.

Carlson, R. W., & Zaki, J. (2018). Good deeds gone bad: Lay theories of altruism and selfishness. *Journal of Experimental Social Psychology*, *75*, 36–40.

Fehr, E., & Fischbacher, U. (2002). Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. *The economic journal*, *112*(478), C1–C33.

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

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, *7*(7), 287–292.

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.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological science*, *26*(5), 633–640.

Jern, A., & Kemp, C. (2014). Reasoning about social choices and social relationships. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, *9*(3), e92160.

McClintock, C. G., & Allison, S. T. (1989). Social value orientation and helping behavior 1. *Journal of Applied Social Psychology*, *19*(4), 353–362.

Murphy, R. O., & Ackermann, K. A. (2011). A review of measurement methods for social preferences. *ETH Zurich Chair of Decision Theory and Behavioral Game Theory Working Paper*.

Rhodes, M. (2013). How two intuitive theories shape the development of social categorization. *Child Development Perspectives*, *7*(1), 12–16.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems* (pp. 1874–1882).

Van Doesum, N. J., Van Lange, D. A., & Van Lange, P. A. (2013). Social mindfulness: skill and will to navigate the social world. *Journal of Personality and Social Psychology*, *105*(1), 86.