Cognition 240 (2023) 105580

Contents lists available at ScienceDirect

Cognition

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

Original articles

Identifying social partners through indirect prosociality: A computational account

Isaac Davis [a,*], Ryan Carlson [a], Yarrow Dunham [a,b], Julian Jara-Ettinger [a,b]

[a] Department of Psychology, Yale University, United States of America
[b] Wu Tsai Institute, Yale University, United States of America

ARTICLE INFO

Keywords:
Theory of mind
Social mindfulness
Computational modeling
Naive utility calculus

ABSTRACT

The ability to identify people who are prosocial, supportive, and mindful of others is critical for choosing social partners. While past work has emphasized the information value of direct social interactions (such as watching someone help or hinder others), social tendencies can also be inferred from indirect evidence, such as how an agent considers others when making personal choices. Here we present a computational model of this capacity, grounded in a Bayesian framework for action understanding. Across four experiments we show that this model captures how people infer social preferences based on how agents act when their choices indirectly impact others (Experiments 1a, 1b, & 1c), and how people infer what an agent knows about others from knowledge of that agent's social preferences (Experiment 2). Critically, people's patterns of inferences could not be explained by simpler alternatives. These findings illuminate how people can discern potential social partners from indirect evidence of their prosociality, thus deepening our understanding of partner detection, and social cognition more broadly.

1. Introduction

A key challenge in selecting social partners is identifying people who are prosocial, cooperative, and mindful of others. Research suggests that people spontaneously detect others' potential value as social partners in terms of their capacity for empathy and prosociality (Fiske, Cuddy, & Glick, 2007; Morelli, Leong, Carlson, Kullar, & Zaki, 2018), but how do people detect these qualities in others? Past work highlights how people draw on *direct* evidence of agents' prosocial inclinations, such as how they respond to requests for aid or social support. Research in this domain suggests that both adults and children can readily infer an agent's prosocial motives from their directly helpful actions (Hamlin & Wynn, 2011; Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Ullman et al., 2009), and use these inferences to form judgments about agents' moral character (Carlson, Bigman, Gray, Ferguson, & Crockett, 2022; Hartman, Blakey, & Gray, 2022; Jara-Ettinger, Schulz, & Tenenbaum, 2020).

In actual social life, however, directly helpful actions are often a consequence of, rather than antecedent to, a social relationship. The person who helps their friend move apartments, for example, may do so because of their close personal relationship to that friend, rather than a general prosocial attitude toward others. In everyday life, it may therefore be difficult to observe direct evidence of prosociality unless one is already a member of (or at least privy to) the social relationship. Furthermore, public displays of directly prosocial behavior may cast doubt on the actor's motives: did Bob agree to help his friend move apartments out of a genuine desire to be helpful, or simply to appear helpful in front of mutual friends (Berman & Silver, 2022)? On the other hand, indirectly prosocial actions are often less confounded by concerns about ulterior motives: the subway rider who offers their seat to a pregnant traveler might do so out of a desire to appear helpful in front of others, but the rider who simply leaves their seat open for the pregnant traveler seems less concerned with their reputation as a helpful person (Siem & Stürmer, 2018).

Given these insights, we propose that an agent's mindfulness of others' welfare when making personal decisions is a potentially powerful cue to their value as a social partner. There is an ample literature showing that observers are sensitive to an agent's awareness of how their actions affect others (Margoni & Surian, 2022; Nobes & Martin, 2022), and how directly prosocial actions influence moral and social evaluations (Heyman, Barner, Heumann, & Schenck, 2014; Poorthuis et al., 2012; Son & Padilla-Walker, 2020). Considerably less work, however, has investigated how indirect displays of prosociality influence social evaluations—the topic we address in this paper.

The decision to act in a way that indirectly benefits others, sometimes called *social mindfulness* (Van Doesum, Van Lange, & Van Lange, 2013; Van Lange & Van Doesum, 2015), can provide strong evidence

* Corresponding author.
E-mail address: isaac.davis@yale.edu (I. Davis).

https://doi.org/10.1016/j.cognition.2023.105580
Received 10 January 2023; Received in revised form 24 July 2023; Accepted 25 July 2023
Available online 10 August 2023

of an agent's prosociality in two ways. First, it shows that the decider is using their theory of mind—the ability to reason about other people in terms of unobservable mental states (Frith & Frith, 2012; Wellman, 2014)—to consider the preferences of others. This is consistent with recent work showing that even children judge social mindfulness by considering whether an agent knows a beneficiary's preferences (Zang, Li, & Zhao, 2023). Second, it signals a willingness to incur a personal cost (e.g.: forgoing a better outcome for oneself) to indirectly benefit someone else (e.g.: enabling them to better fulfill their own preferences) (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Fehr & Fischbacher, 2002; McClintock & Allison, 1989; Van Doesum et al., 2013; Van Lange, 1999). People who possess both of these qualities (those who are socially mindful) are more likely to care for and effectively support others, making them especially valuable social partners (Van Doesum et al., 2013; Van Doesum et al., 2020). Thus, the capacity to detect prosociality from indirect evidence may be crucial for navigating human social life.

How do people detect socially mindful agents? Previous research suggests that, in social interactions, people interpret each others' behavior by inferring the mental states which best explain that behavior (Dennett, 1989; Gergely & Csibra, 2003a). In particular, inferences about agents' mental states are structured around an assumption that agents act to maximize their subjective utilities—the difference between the rewards they receive and the costs they incur (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jern, Lucas, & Kemp, 2017; Lucas et al., 2014). Importantly, such models can also express social preferences in terms of recursive utilities, i.e.: the value an agent assigns to the outcomes of others (Jern & Kemp, 2014; Ullman et al., 2009). However, it remains unclear how people identify social partners from indirect (rather than direct) displays of prosociality.

Our work has two goals. First, we present a computational account of how people assess agents' social preferences from indirect evidence. We focus on settings where agents' choices can reveal whether they considered how their actions indirectly affect others, and how they value others' outcomes when making these choices. This model extends prior work (Jern & Kemp, 2014) to highlight the role of an agent's perceived knowledge or ignorance about the preferences of others, and how this shapes observers' assessments of the decider's prosociality and theory of mind.

Our second goal is to test our computational model in four experiments focusing on subtle displays of social mindfulness in sequential resource-allocation tasks. Scenarios like these frequently occur in everyday social life. Whether choosing how much space to take up on a crowded bus, or deciding how long to utilize shared resources at the gym, people often make choices that will impact which options remain for others. Thus, choices like these can reveal whether the decider deployed their theory of mind to try to indirectly benefit others, thereby providing indirect evidence of the decider's prosociality. In Experiments 1a through 1c, we investigated how people interpret an agent's personal choices in light of what the agent preferred, and crucially, what the agent believed others preferred (Zang et al., 2023). Our results demonstrate that people's judgments about the decider's social preferences (the degree to which the decider values others' utility relative to their own) closely track with the predictions of a utility based model, and that simpler models which ignore the decider's knowledge of others fail to explain these judgments. In Experiment 2, we demonstrate that the inverse is also true. That is, when a decider's social preferences are known (for instance, that the agent is known to care about others), observers use an agent's choices to infer whether they knew the preferences of others, and what they believed those preferences were. This demonstrates not only that people can incorporate evidence that an agent is using theory of mind into their inferences of that agent's social preferences, but also that people can infer what agents know or believe about each other based on their social behavior. These experiments provide converging support for a utility based model of how people identify prosocial partners from indirect evidence.

2. Computational framework

Our framework models scenarios in which an agent A ("actor") makes a decision that indirectly affects another agent B ("bystander"). Importantly, the identity of the bystander does not need to be known to A when making the choice (and thus, B can represent a specific agent, or it can be a general representation of the fact that some unknown agent might be affected). For our experiments, we used a particular form of decision problem that allows agents to be indirectly prosocial (Van Doesum et al., 2013; Zhao, Zhao, Gweon, & Kushnir, 2021). In these tasks, A is given first pick from a common supply of items of two types (e.g.: snacks of two different types), and permitted to take a fixed number of those items. A knows that B will pick second, and that B will therefore have access to whatever items A does not take. Based on A's decision, we prompted participants to infer either the degree to which A cares about B (i.e. A's "social preference"; Experiment 1), or A's beliefs about B's preferences for the available items (i.e. A's "social knowledge"; Experiment 2).

We modeled participants' inferences in both experiments using a single framework built on two core assumptions. The first is that social inference is structured around the expectation that agents act rationally to maximize their utilities—the differences between the costs they incur and the rewards they obtain (Gergely & Csibra, 2003b; Jara-Ettinger et al., 2016; Jern et al., 2017; Lucas et al., 2014). We encoded this assumption into a generative model of agent behavior (Section 2.1). Our second assumption is that observers can invert this generative model to infer the underlying utilities and/or beliefs that best explain an observed action, and we implement this inversion as a form of Bayesian inference over the aforementioned generative model (Section 2.2). Previous work has leveraged similar assumptions to explain a range of social inferences about goals (Baker, Saxe, & Tenenbaum, 2009), preferences (Jern et al., 2017; Lucas et al., 2014), beliefs (Baker et al., 2017; Tauber & Steyvers, 2011), competence (Jara-Ettinger et al., 2020), and social intentions (Ullman et al., 2009).

2.1. Generative model

A defining feature of our generative model is that A may care not only about how their decision affects them, but also how their decision affects others. We capture this notion using a recursive utility function. First, we assume each agent has an individual reward function R, which captures the reward that the agent derives from an allotment of items. For example, if there are two types of treats (say, donuts and cupcakes), then $R_A([1,1])$ is the total reward that A derives from having one donut and one cupcake, while $R_B([0,2])$ is the total reward B derives from having zero donuts and two cupcakes. For clarity, we first discuss how our model uses reward functions to make decisions, and then turn to discuss the internal structure of these reward functions (Section 2.1.1). To simplify notation, we will use d to denote A's decision (i.e.: which items to keep), and $R_A(d)$ and $R_B(d)$ to denote the respective rewards that A and B receive from this decision (i.e.: $R_A(d)$ is A's reward from the items they keep, while $R_B(d)$ is B's reward from the items that remain after A's decision).

Given these individual reward functions, we assume that the total utility $U(d)$ which A derives from the outcome of decision d is a weighted sum of A's personal reward and B's personal reward (following Jern & Kemp, 2014; Powell, 2021), i.e.:

$$U(d) = w_A R_A(d) + w_B R_B(d) \tag{2.1}$$

The pair of weights $W = (w_A, w_B)$ denotes A's social preference, where $0 \leq w_A \leq 1$ captures A's *egocentric* preference and $0 \leq w_B \leq 1$ captures A's *allocentric* preference towards B (Murphy & Ackermann, 2011).[1] For example, $W = (1,0)$ represents a case where A only values

[1] This formulation also allows using negative weights to capture A's "dislike" for B, though our experiments focus only on self-interest versus other-interest, and do not invoke negative social preferences.

their own reward and is therefore completely self interested, while $W = (0, 1)$ reflects a completely altruistic A who only values B's reward. Depending on the context, these weights may reflect either A's social preference towards a particular B, or A's general social preference towards others, and we test both interpretations experimentally. In either case, Eq. (2.1) reflects the total utility that A receives from decision d, including any utility A receives indirectly from B's reward.

Note that Eq. (2.1) assumes A already knows B's reward function. In reality, however, we often face varying degrees of uncertainty about the preferences of others: we may know what our close friends and family like or dislike, but have no idea about the preferences of strangers, or have only a vague idea about the preferences of a recent acquaintance (e.g.: A might think B prefers cupcakes over donuts, but not know the strength of the preference). To capture this uncertainty, we model A's beliefs about B as a probability distribution $K(R_B)$ over possible reward functions. This distribution K reflects the degree of A's knowledge about B's preferences: if A knows B's preferences exactly, then K is a point mass distribution centered on the correct reward function. If A has no knowledge whatsoever, then K is a uniform distribution over all possible reward functions. Importantly, K can also encode partial knowledge: for example, if A believes that B prefers cupcakes over donuts, but does not know the strength of this preference, then K is a distribution over all reward functions with the property that $R_B(\text{cupcakes}) > R_B(\text{donuts})$.

Given this belief K, A's total *expected* utility from decision d is given by

$$U(d) = w_A R_A(d) + w_B \mathbb{E}[R_B(d)] \tag{2.2}$$

where the expectation is taken with respect to A's belief distribution $K(R_B)$. That is, $\mathbb{E}[R_B(d)] = \int x P(x) dx$, where $P(x)$ is the subjective probability that $R_B(d) = x$. Given this utility function, we assume that A will make decision d using a softmax decision function, so that A is more likely to make decisions that yield a higher total expected utility. The parameter value for this softmax function was tuned in an earlier pilot study (see Supplemental Materials).

2.1.1. Reward functions

In all of our experiments, each trial depicted a scenario in which A is given first pick from a common supply of food items of two types, and allowed to keep a fixed number of those items. A knows that B will pick second, and that B will therefore receive whatever items A does *not* take. To compute the total reward that each agent receives directly from the items they take, we assumed that each agent has reward values r_1 and r_2 for each item respectively. For example, r_1^A denotes the reward that agent A receives from one unit of item 1, and r_2^B denotes the reward that agent B receives from one unit of item 2.

In an initial version of our model, we computed each agent's final reward as the sum of the rewards of each individual item they obtain. For example, if agent A receives allotment $[X, Y]$ (i.e.: X of item 1 and Y of item 2), their final reward is the sum $R_A([X, Y]) = r_1^A * X + r_2^A * Y$. We soon discovered, however, that when A does not know B's preferences (which is especially critical when B is an unknown agent who will arrive at a later time), this linear reward function means that any action that A takes is equally socially mindful. This is because, mathematically, any combination of items will have the same expected reward for B (e.g., the expected reward for two items of category 1 is the same as the expected reward for two items of category 2, which is the same as the expected reward for one item from each category). However, this conflicted with one of the widely document empirical findings that motivated our model in the first place: when there is uncertainty about another agent's preferences, leaving that agent a choice of items (e.g.: one of each) is interpreted as more "mindful" than leaving the agent two of one item.

In response to this, we reasoned that the rewards may have a high rate of saturation (e.g.: one's desire for additional cupcakes may decrease sharply after having a first cupcake), and we therefore hypothesized that observers assume some degree of discounting in the individual reward functions (Baucells & Sarin, 2010). To this end, we introduced a discount parameter $0 < \rho < 1$, so that if A receives a reward r from eating one cupcake, A will receive $\rho * r$ from the second cupcake, $\rho^2 * r$ for the third, and so on. We then computed the reward that the agent receives from allotment $[X, Y]$ as

$$R_A([X, Y]) = r_1^A * D(X, \rho) + r_2^A * D(Y, \rho) \tag{2.3}$$

where $D(X, \rho)$ applies the discount rate function described above. Rather than attempting to estimate the discount rate from human data (as we expected it to vary across participants), we instead integrated this parameter out of the model using a $Beta(\alpha, \beta)$ prior distribution, and tuned the hyper-parameters α, β using an initial pilot study (See supplemental materials). This is equivalent to modeling social inference under the assumption that agents derive diminishing marginal utilities from additional food items, but without knowing each agent's true discount rate. As described further in the next section, we validated our discounting assumption by comparing our results against an alternative model with no discounting (i.e.: fixing $\rho = 1$).

Note that this observation—linear reward functions imply that any choice is equally socially mindful—is not visible in paradigms where the second agent can choose only one item from the ones that the first agent left behind, and only appears when the second agent will be allowed to take *all* of the items left behind. Therefore, the inclusion or absence of discounting does not change the model's predictions in the traditional version of the task: when agent B only gets to pick one item, their total utility is just the utility they receive from that item, and discounting does not apply.

2.2. Inference

The generative model described above captures the proposed intuitive causal model that people use to reason about each other's behavior. With this model, people can make predictions about how A will behave—expressed as a probability distribution $P(d|R_A, K, W)$—as a function of A's personal preferences R_A, social preferences W, and beliefs K about the other agent's preferences. Equipped with this model, an observer can leverage this distribution to infer an agent's mental states using Bayesian inference.

Our experiments focused on two types of inference. In Experiment 1, participants observed A's personal preferences (R_A), A's beliefs about B's preferences (K), and A's decision d, and were then asked to infer A's social preference $W = (w_A, w_B)$. In this situation, the observer can infer A's social preference according to

$$P(W|d, R_A, K) \propto P(d|R_A, K, W) P(W|R_A, K) \tag{2.4}$$

In other words, the probability that A's social preference is equal to W (given A's personal preference, belief, and decision), is proportional to the probability that A would choose d (given A's social preference) times the prior probability of A's social preference.

In Experiment 2, participants observed A's personal preferences, A's social preference, and A's decision, and were then asked to infer (a) whether A knows which item B prefers and (b) which item A thinks B prefers (assuming that A *does* know B's preference). In this case, inferences about K can be captured by

$$P(K|d, R_A, W) \propto P(d|R_A, K, W) P(K|R_A, W) \tag{2.5}$$

When computing the prior distribution over A's knowledge $P(K|R_A, W)$, we encoded an assumption that A is more likely to attend to and remember B's preference the more A cares about B. This is based on the idea that, in order to fulfill another person's preferences, one must know what those preferences are in the first place. Thus, the more A cares about B (and therefore values fulfilling B's preferences), the more motivation A has to attend to and remember B's preferences. Note

that in some settings, selfish and manipulative people may also benefit from attending to others' preferences, so as to better leverage that knowledge when manipulating others. In our task, however, there is no possible way for people to "manipulate" each other, and we therefore only considered agents who range from being apathetic (i.e., not caring about others) to socially mindful (i.e., positively caring about others). Our choice of a prior is therefore consistent with both the possibility that social knowledge is correlated with social mindfulness, and with the possibility that social knowledge is correlated with extreme social attitudes (either positive social mindfulness or negative antagonistic attitudes).

To this end, we define the prior $P(K|R_A, W)$ so that the probability that A is totally ignorant of B's preferences is proportional to $w_A/(w_A + w_B)$ (See supplemental materials for the exact form of this prior). Note that this prior is only relevant for generating Experiment 2 predictions, as the term $P(K|R_A, W)$ does not appear in the equation for Experiment 1 (since A's beliefs are fixed in the Experiment 1 tasks). We therefore used Eqs. (2.4) and (2.5) to generate predictions for Experiments 1 and 2, respectively.

3. Experiments

In this section we present two types of model evaluations. First, we compared our model predictions against participant judgments of A's social preference, based on how A's decisions indirectly affect others (e.g., inferring that A is prosocial if they forego their favorite treat so that others may have it; Experiments 1a, 1b, & 1c). Second, we compared our model predictions against participant judgments about A's social knowledge, based on A's social preferences and decisions (e.g., if we believe A to be mindful of others and observe A taking the last scone, this might suggest that A believes no one else likes scones; Experiment 2).[2]

3.1. Experiment 1a

3.1.1. Participants

40 adult participants with US-based IP addresses were recruited via Amazon Mechanical Turk (mean age = 40.1, S.D. = 11.8). Four additional participants were recruited but excluded for failing at least one comprehension check question on each of two attempts. (See supplemental materials for survey questions and comprehension checks).

3.1.2. Stimuli

Stimuli consisted of 24 trials (see Fig. 1 for examples). Each trial contained a pictogram showing agents A and B, each with randomly chosen names, and a box containing brownies and cupcakes. Depending on the trial, the box contained either two of each treat (four total), or one of one type and two of the other (three total). The top of the screen showed a description explaining that A had to share two of the treats with B. A's material preferences were depicted using a verbal description in a thought bubble (either "I prefer [brownies/cupcakes]" or "I like both treats equally"), as were A's beliefs about B's material preferences (either "B prefers [brownies/cupcakes]", "B likes both treats equally", or "I don't know what B prefers"). Finally, the pictogram showed a box containing whichever treats remained after A made their choice (either two brownies, two cupcakes, or one of each). Altogether, this created a combinatorial space varying the number of initial treats (3 or 4), A's personal preference, A's knowledge (or lack thereof) of B's personal preference, and A's choice. This yielded 6 unique "3-treat" trials, and 18 unique "4-treat" trials, for a total of 24 trials (see Supplemental Materials for the full combinatorial space of trials).

3.1.3. Procedure

Participants first read a cover story in which one agent (A) can take one (in the 3-treat case) or two (in the 4-treat case) treats from a box of cupcakes and brownies. Participants were told that A knows B will get whichever treats A leaves behind, and were then taught how to interpret the information in each stimulus. Participants were then given a 6-question multiple-choice comprehension check (see Supplemental materials). Participants who failed at least one question on the first try were shown the instructions again, followed by a second attempt at the comprehension check questions. Those who also failed at least one question on the second attempt were excluded from the study. Participants that qualified for the task were shown all 24 trials in a random order upon completing the comprehension checks. For each trial, participants were asked "how much do you think A values [his/her] own preferences" and "how much do you think A values B's preferences", and gave each answer using a continuous sliding scale from 0 ("not at all") to 5 ("a lot").

3.1.4. Model predictions and alternate models

We generated model predictions using Eq. (2.4), which takes A's personal reward function R_A, A's belief K about B's preferences, and A's decision d, and outputs a probability distribution $P(W|d, R_A, K)$ over A's social preference W. To implement this, we needed to transform the verbal belief description in each stimulus (e.g.: "I think B prefers cupcakes over brownies", or "I don't know what B prefers") into a probability distribution K over possible reward functions. If A is totally ignorant of B's preferences, we treat K as a uniform distribution over all possible reward functions. If A believes that B prefers cupcakes, we condition this uniform distribution on the observation that R_B(cupcake) > R_B(brownie) (i.e.: that B gets more reward from cupcakes than brownies) using Bayesian posterior updating, and similarly if B prefers brownies (see Supplemental Materials for full details). Following this procedure, we used Eq. (2.4) to compute the distribution $P(W|d, R_A, K)$ for each trial, then used the expected value of this distribution as our prediction of A's social preference.

To better understand our model predictions, we also implemented three alternative models, each designed to evaluate a different aspect of our main account. First, our computational model assumes that rewards are discounted (i.e., the more brownies you eat, the less enjoyable they become). We therefore implemented a "no-discounting" model, which is identical to the main model, with the difference that the rewards are not discounted.

Our last two models were designed to explore alternative accounts that do not require the full Theory of Mind reasoning present in our main model. One possibility is that people infer social preferences in terms of how much an agent sacrifices themselves. For example, if A prefers brownies over donuts, giving both brownies to B would reveal a large "sacrifice" in A's personal reward, regardless of B's actual preference for brownies. Such an account would only require thinking about the difference between A's highest possible reward and the true reward they obtained, without considering what A believes about B's preferences (i.e.: the greater A's sacrifice, the more prosocial they are). We call this model the "self-sacrifice" model.

Conversely, it is possible that people assume that, the larger of a reward B gets, the more pro-social A is, regardless of A's own reward. We therefore implemented a simple "other-outcome" model that equates B's rewards with A's social tendencies (i.e.: the greater B's reward, the more prosocial A is). Classic literature on prosociality and altruism suggests that both of these factors are potential cues to an agent's prosociality (Swap, 1991), and we included these models to evaluate whether either of these features alone is sufficient to capture participant inferences.

3.1.5. Results

For each of the 24 trials, participants provided separate judgments about self-weight (w_A in the model) and other-weight (w_B in the model), yielding 48 total inferences. Per our pre-registered analysis plan (see osf.io/23rq7) we Z-scored responses within participants and across trials (separately for each parameter), then averaged across

[2] Data and analysis scripts for all experiments are available at osf.io/8extq/

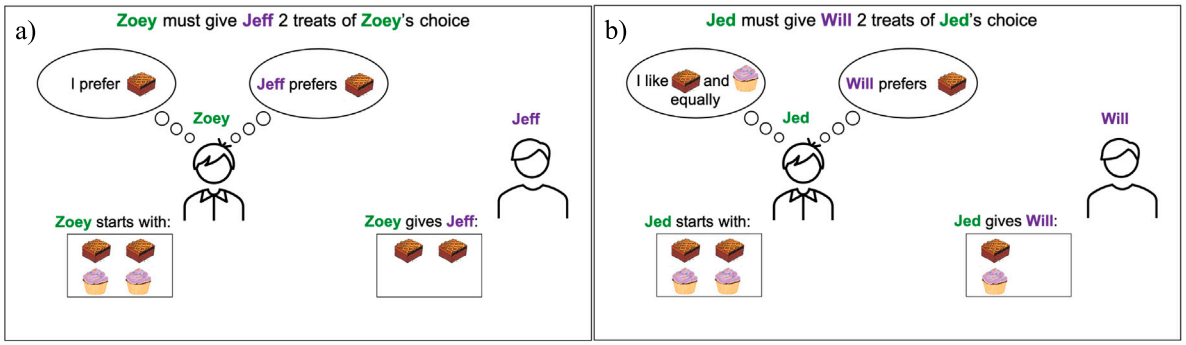

Fig. 1. Two examples of stimuli from Experiment 1a. Agent A's material preferences and beliefs about B's preferences are shown in the two thought bubbles.

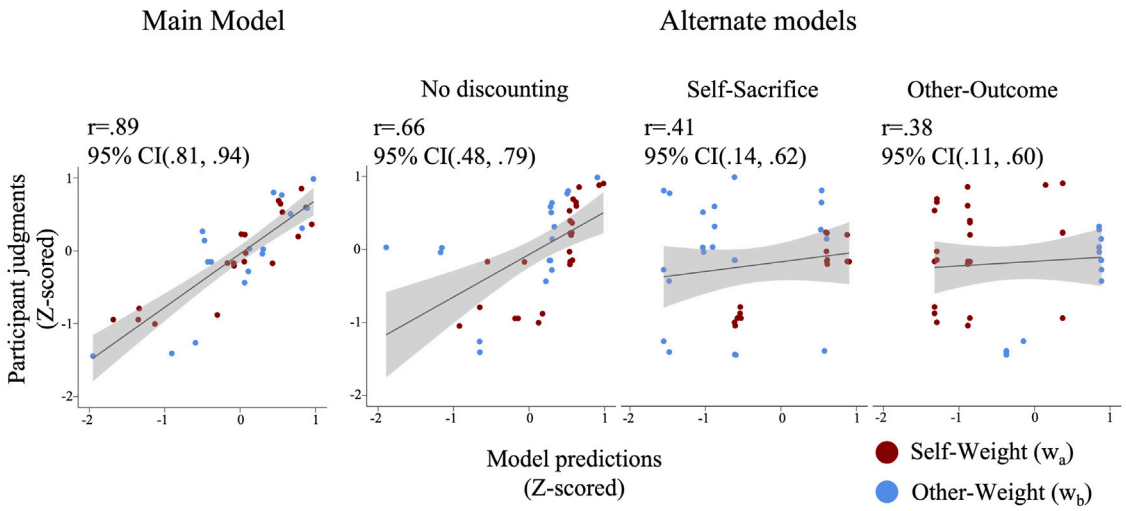

Fig. 2. Results from Experiment 1a. Each subplot shows a comparison between predictions generated by one of the four models (including the three alternate models) against participant responses. Each point represents a judgment from a single trial: red dots correspond to self-weight (ω_a) and blue dots correspond to other-weight (ω_b). The x-axis depicts model predictions (z-scored) and the y-axis depicts participant responses (z-scored within participant and averaged across participants). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

participants for each trial. We then Z-scored the mean parameter values outputted by the model, and computed Pearson correlations with participant judgments. As shown in Fig. 2, main model predictions were highly and significantly correlated with participant responses (r = .89, 95% CI(.81, .94)), and were significantly more correlated with participant responses than all three alternate models (defined as a bootstrapped difference in correlations with 95% CI not crossing 0—see Supplemental Materials).

Fig. 3 shows the results for each trial in Experiment 1a. We highlight two pairs of trials that illustrate our model's qualitative behavior and how it matched human intuitions. These cases all show events where A prefers brownies over cupcakes, and encounters a box with two brownies and two cupcakes. In the first pair (second row of Fig. 3, second and third column from the left, highlighted in red), A did not know B's preference. In the left trial, where A took a brownie and a cupcake (leaving B with a brownie and a cupcake), both participants and our model inferred that A was selfless and cared about B. By contrast, when A took both brownies (leaving B with two cupcakes), our model and participants both inferred that A cared highly about their own rewards and less so about others. This is consistent with previous findings on the "social mindfulness" phenomenon (Van Doesum et al., 2013; Zhao et al., 2021), whereby people associate prosocial attitudes with actions that preserve the greatest number of options for others (in this case, leaving one of each treat for the next person).

However, the next two cases (third row of Fig. 3, second and third column from the left, highlighted in blue) show how preserving choice for others is not always considered the most socially mindful action. These two trials show identical scenarios to the first two, with the difference that A knew that B preferred cupcakes. In the left trial, when A made the "option-maximizing" decision (leaving one of each for B), participants and the model both inferred that A did not care about B. However, when A kept both brownies, participants and the model inferred that A cared about both B's preferences and their own preferences, as this action maximized both agents' rewards without conflict. Thus, our results suggest that the classic social mindfulness effect is a special case of this more general utility based reasoning, and that the effect only holds when A does not know B's preference.

Fig. 4 shows three trials that highlight the types of cases where the alternative models failed to explain participant judgments. In the first example, A liked both treats equally and did not know B's preference. When A took both brownies, participants and the main model both inferred that A did not care about B, while the no-discounting model inferred that A was prosocial. Intuitively, this is because, when A is totally uncertain about B's preferences, B is equally likely to prefer either treat. In the absence of discounting, it follows that B is equally likely to enjoy any combination of two treats: two brownies, two cupcakes, or one of each. Since all three choices provide the same expected reward to B, A's choice does not reveal how much they care about B's rewards. When the rewards are discounted, however, the second treat of the same type contributes a diminished amount to B's total rewards, so B is more likely to prefer one of each rather than two of the same type. Thus, when A is uncertain about B's preference, A's

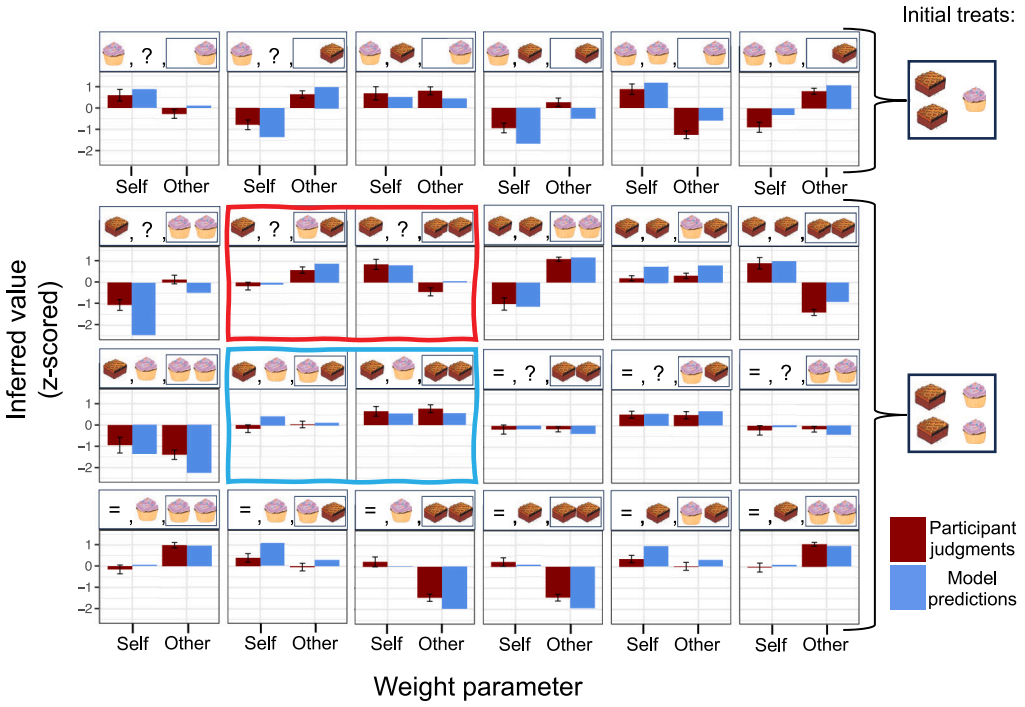

Fig. 3. Model predictions and mean participant judgments for all 24 trials in Experiment 1a. Trial configurations are shown at the top of each chart. Reading left to right, the headings indicate (1) A's personal preference, (2) A's belief about B's personal preference, and (3) which treats A chose to keep. Bar charts show participant judgments (z-scored within participants and averaged across participants) compared against z-scored model predictions. Vertical bars show bootstrapped 95% confidence intervals. The two pairs of highlighted trials demonstrate the "social mindfulness" effect as described in Section 3.1.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

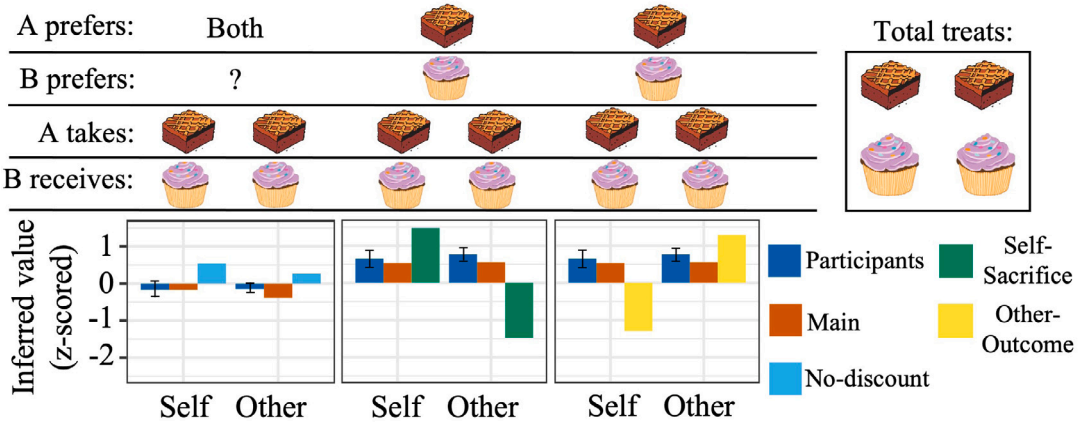

Fig. 4. Three trials from Experiment 1a demonstrating the qualitative failures of each alternate model, respectively. Each column shows a single trial and a different alternate model. Rows depict (from top to bottom): which type of treat A preferred, which type of treat A believed that B preferred (question marks indicate that A did not know), which two treats A took from the box, and which two treats were left for B after A's choice. Each bar chart shows z-scored predictions from one of the three alternate models compared against participant judgments (z-scored within participants and averaged across participants) and z-scored predictions from the main model. Vertical bars show bootstrapped 95% confidence intervals.

choice only reveals A's social preference if we assume that rewards are discounted.

The next trial (middle column; Fig. 4) shows how the "self-sacrifice" model failed to explain participant intuitions. In this case, A preferred brownies and knew that B preferred cupcakes, so when A kept both brownies and left both cupcakes, this decision maximized both agents' rewards without conflict. Participants and the main model therefore inferred that A likely cared about both their own reward and B's reward. The self-sacrifice model, however, only reflects that A's decision maximized their own personal reward, which means that A did not sacrifice anything with this decision. The self-sacrifice model therefore inferred that A is very selfish (i.e., $w_A \gg w_B$).

Finally, the last trial illustrates how the "other-outcome" model failed to explain participant intuitions. In this case, A preferred brownies and knew that B preferred cupcakes, so when A kept both brownies and left both cupcakes, this decision maximized both agents' rewards without conflict. Participants and the main model therefore inferred that A likely cared about both their own reward and B's reward about equally. However, because the outcome model focuses only on the total reward that B obtains, it inferred that A cared much more about B's reward than their own. That is, despite correctly identifying that A's decision maximized B's reward, this model failed to account for the fact that A's decision also maximized their own reward. Thus, these results demonstrate (a) that participants attended to both the costs A incurred

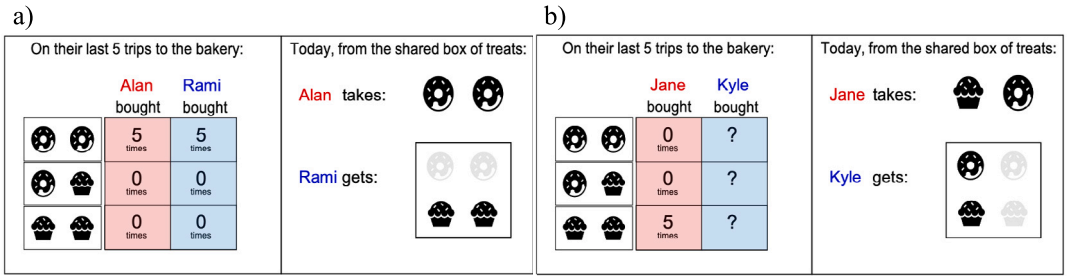

Fig. 5. Two examples of stimuli from Experiment 1b. Agent A's material preferences and beliefs about B's preferences are depicted as numerical choice histories. Question marks indicate that A does not know the choice history of that agent.

and the expected rewards that B received, and (b) that participant inferences reflected an assumption of discounted utilities, particularly in cases where A does not know B's preference.

3.2. Experiment 1b

Experiment 1a showed that people can infer social preferences by considering how an agent's personal choices affect others. However, this experiment had two limitations. First, in Experiment 1a, the stimuli always read "A must give B 2 treats of A's choice". This wording therefore highlighted that A was directly making a choice that would affect both members. It is therefore possible that our model fit participant judgments only because participants were explicitly told that A was forced to choose which rewards B would get, which conflicts with our primary aim of evaluating social inference from indirect evidence of prosociality. To address this concern, we modified the cover story in Experiment 1b to a situation where A was clearly making a personal choice that indirectly affected B.

A second limitation of Experiment 1a is that it revealed people's preferences using qualitative descriptions, which restricted us to only three preference categories (preferring one treat or the other, or liking both equally). To test whether people can make these inferences from more granular information, the stimuli for Experiment 1b depicted preferences using quantitative choice histories, rather than verbal descriptions

3.2.1. Participants

40 adult participants with US-based IP addresses were recruited via Amazon Mechanical Turk (mean age = 40.2, S.D. = 10.3). 6 additional participants were recruited but excluded for failing at least one comprehension check question on each of two attempts.

3.2.2. Stimuli

The stimuli for Experiment 1b depicted the same variables as 1a. The key difference was that A's personal preferences and social knowledge were depicted indirectly, using numerical choice histories, rather than direct verbal descriptions. To this end, participants were told that the agents often visit the same bakery to buy themselves treats, and will sometimes observe what other agents buy for themselves. In lieu of verbal preference and belief descriptions, each trial displayed A's choice history over the previous 5 visits to the bakery, and what A had observed B buy for themselves over their previous 5 visits to the bakery (or a column of question marks, indicating that A did not visit the bakery at the same time as B and therefore did not see what B purchased). This enabled us to represent clear preferences as we did in Experiment 1a (e.g.: "B prefers brownies" could be reflected by B choosing to buy brownies in all 5 trials), but also enabled us to represent ambiguous or mixed preferences (e.g.: choosing brownies 3 times and donuts twice), which we could not represent in the stimuli for 1a. Fig. 5 shows examples of stimuli from this experiment.

The combinatorial space for Experiment 1b was significantly larger than that of 1a: each choice history consisted of 5 decisions divided between three possible choices, yielding 21 possible choice histories for A, 22 for B (including the "all question marks" history), and $21 \times 22 = 462$ possible pairs of choice histories. Combining this with A's three possible decisions yielded 1386 possible trials. To select 25 trials that reflected a broad distribution of social preferences, we generated model predictions for all 1386 conditions, and computed, for each condition, the ratio of inferred weights $w_A/(w_A + w_B)$, where a ratio of 1 denotes a purely selfish A, a ratio of 0 entails a purely altruistic A, and a ratio of .5 entails a fully egalitarian A. We then defined 5 social preference categories (very selfish, weakly selfish, roughly egalitarian, weakly altruistic, and very altruistic), and for each category, selected the 5 trials for which the model's predictions most closely fit that category. For example, the "very selfish" category contained the 5 trials with the highest weight ratios, and the "very altruistic" category contained the 5 trials with the lowest weight ratios. This yielded 25 total trials, with 5 for each social preference category.[3]

3.2.3. Procedure

Experiment 1b was similar to Experiment 1a with a modified cover story: participants were introduced to an office setting where a manager had left a box of four snacks for two employees to enjoy. Participants were told that one agent (A) worked the morning shift, while the other (B) worked the afternoon shift, so that A would get to choose their two snacks first. Participants were additionally told that A knew that B works the afternoon shift.

Participants were then taught how to interpret the stimuli (Fig. 5) and then given two chances to pass a 6 question comprehension check ensuring that they were paying attention and could correctly interpret the stimuli (see Supplemental materials for details). Participants who failed at least one question on the first attempt were shown the instructions a second time, followed by a second attempt at the comprehension checks. Participants who failed one or more questions on the second attempt were excluded from the study. Participants who passed the questionnaire were then shown all 25 trials in a random order. For each trial, participants are asked to rate (a) how much A cares about him/herself and (b) how much A cares about B. Both responses were entered on a continuous sliding scale from 0 (does not care at all) to 5 (cares a lot).

3.2.4. Model predictions and alternative models

Model predictions for Experiment 1b were generated using the same procedure as Experiment 1a (see Section 2.2 for full details). To implement the model, we needed to transform the numerical choice histories in the stimuli into probability distributions over possible reward functions. To this end, we transformed the "full ignorance" condition (in which A has no evidence of B's preferences) into a uniform distribution over reward functions. The distribution for each non-ignorance condition was then computed by conditioning this uniform prior on

[3] See Supplemental Materials for a full list of trials and category explanations.

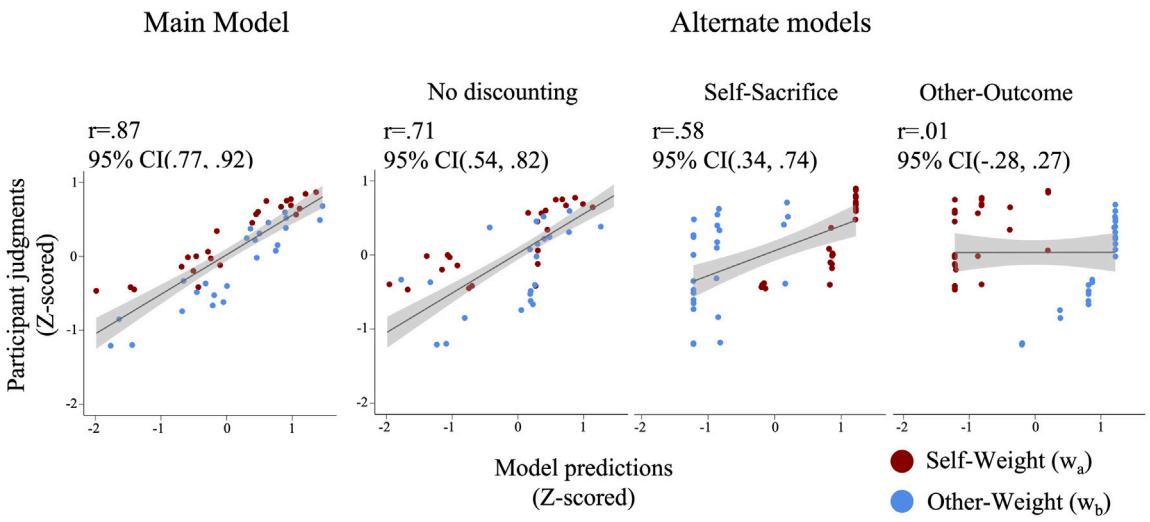

Fig. 6. Results from Experiment 1b. Each subplot shows a comparison between predictions generated by one of the four models (including the three alternate models) against participant responses. Each point represents a judgment from a single trial: red dots correspond to self-weight (w_a) and blue dots correspond to other-weight (w_b). The *x*-axis depicts model predictions (z-scored) and the *y*-axis depicts participant responses (z-scored within participant and averaged across participants). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the agent's observed choices, using Bayesian posterior updating (see Supplemental Materials for full details). Finally, we generated three additional sets of predictions using the same three alternate models described in Experiment 1a ("no discounting", "self-sacrifice", and "other-outcome").

3.2.5. Results

Data were analyzed in the same way as Experiment 1a, in accordance with our pre-registration (see osf.io/pn6zs). As shown in Fig. 6, our model closely tracked average participant responses, with correlations similar to the results of Experiment 1a (r = .87, 95%CI (.77, .92)). This reinforces our findings from Experiment 1a, and further suggests that people can integrate graded, quantitative information about preferences into their social inferences. Furthermore, whereas the stimuli from Experiment 1a depicted A giving two treats to B, the cover story and stimuli from Experiment 1b made it clear that A chose which treats to keep for themselves, rather choosing which treats to give to B. This demonstrates that people make similar inferences when A's decision only affects B's outcome indirectly.

Also similar to Experiment 1a, a bootstrapped difference in correlations revealed that all three alternate models were significantly less correlated with participant data than the main model, ("no discounting": r = .71, 95%CI (.54, .82); "self-sacrifice": r = .58, 95%CI (.34, .74); "other-outcome": r = .01, 95%CI (-.28, .27)). These results further support that participant judgments reflected a discounted expected utility computation (at least in cases where B's preferences suggested that B's reward functions were in fact discounted), and were sensitive to both the cost that A incurred and the benefit that A indirectly yielded to B.

3.3. Experiment 1c

While the previous two experiments provide evidence for our utility-based account of social preference judgments, the cover stories for those experiments leave some potential ambiguities when interpreting their results. First, because each stimulus depicted two particular agents (A and B), it is unclear whether participant responses reflected their judgments about A's relationship with B specifically, or A's general disposition towards others. It might be the case, for example, that A acts generously towards their friend B, but selfishly towards the rest of their co-workers. However, evaluating a potential social partner typically requires assessing the agent's general disposition towards others, which is not necessarily reflected by their interactions with one specific person. Additionally, the fact that A knows who will get the next choice makes it ambiguous as to whether participants really interpret these interactions as "indirect" displays of prosociality. To resolve these ambiguities, we replicated Experiment 1a with two key changes to the cover story and stimulus information: first, the decider agent A no longer knew who would get next pick from the box of snacks. Second, A's belief about the other agent's preferences was depicted as a general belief about the preference distribution of A's coworkers (e.g.: "my co-workers usually prefer X over Y") rather than beliefs about a specific co-worker (e.g.: "B prefers X over Y").

3.3.1. Participants

40 adult participants with US-based IP addresses were recruited via Amazon Mechanical Turk (mean age = 36.8, S.D. = 10.1). 4 additional participants were recruited but excluded for failing at least one comprehension check question on each of two attempts.

3.3.2. Stimuli

Experiment 1c depicted the same variables as 1a, and used the same set of 24 trial configurations. The key difference was that the Experiment 1c stimuli only depicted the decider agent A, and did not specify which other agent would get to pick after A. Similarly, rather than depicting A's beliefs about the preferences of another specific agent, the stimuli depicted A's beliefs about the general distribution of their co-workers' preferences (either "my coworkers usually prefer [brownies/cupcakes]" or "I'm not sure what my coworkers prefer"). Finally, we explicitly showed both agent A's choice and the treats that remained after A made their choice, despite the fact that one could be easily inferred from the other, so as to minimize the possibility that a participant might confuse the two. Fig. 7 shows two examples of stimuli from Experiment 1c.

3.3.3. Procedure

The procedure for Experiment 1c was nearly identical to Experiment 1a, with one minor alteration to the cover story. In particular, we specified that agent A knew that only one other agent would be working that day (and would therefore get to keep whichever treats A left behind), but A did not know which agent it would be. Other than this change, the instructions, comprehension check, and procedure for Experiment 1c were identical to Experiment 1a.

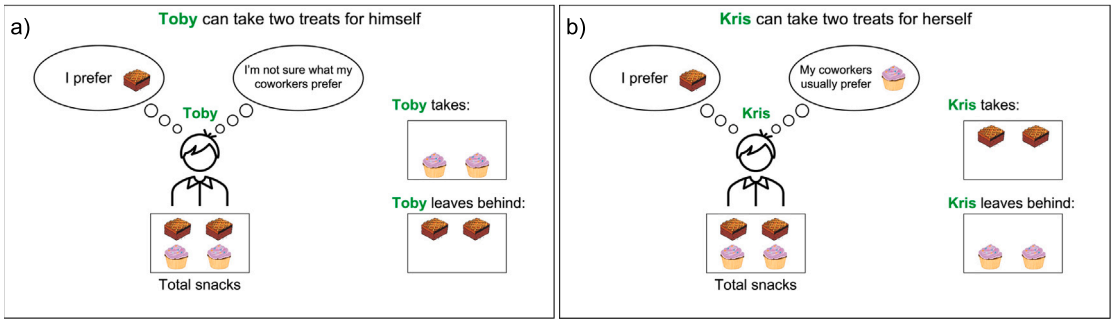

Fig. 7. Two examples of stimuli from Experiment 1c. Agent A's material preferences and beliefs about their coworkers' preferences are shown in the two thought bubbles.

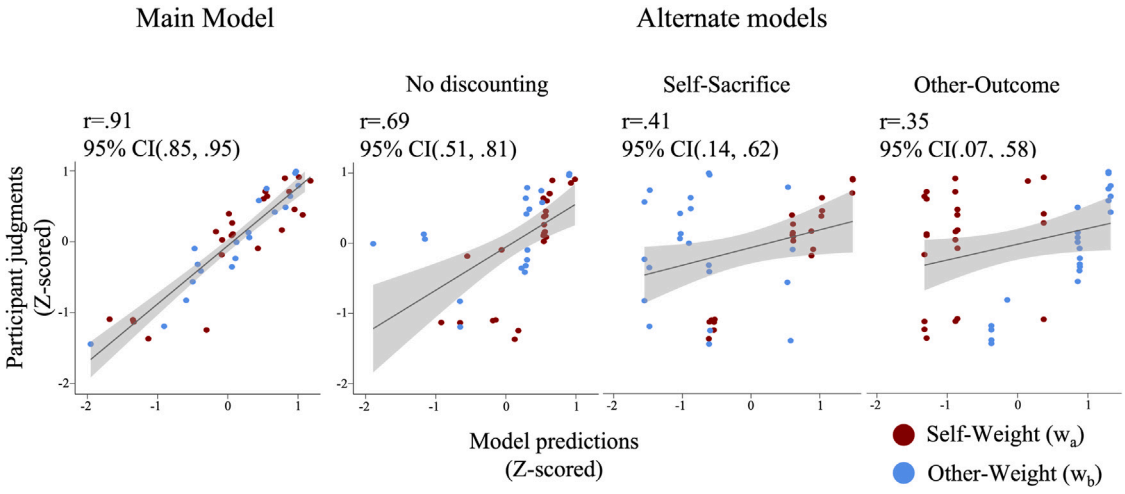

Fig. 8. Results from Experiment 1c. Each subplot shows a comparison between predictions generated by one of the four models (including the three alternate models) against participant responses. Each point represents a judgment from a single trial: red dots correspond to self-weight (w_a) and blue dots correspond to other-weight (w_b). The x-axis depicts model predictions (z-scored) and the y-axis depicts participant responses (z-scored within participant and averaged across participants). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3.4. Model predictions and alternate models

We used the same set of alternate models and the same procedure for generating all model predictions that we used in Experiment 1a. Because we used the exact same set of trial configurations, all predictions were identical with those from Experiment 1a.

3.3.5. Results

As pre-registered (osf.io/cqwev), we analyzed the data from Experiment 1c following the same procedure used for Experiment 1a. As shown in Fig. 8, the results were also very similar to Experiment 1a: our main model closely tracked average participant responses (r = .91, 95%CI (.85, .95)), while our three alternate models were significantly and substantially less correlated with participant responses ("no discounting": r = .69, 95%CI (.51, .81); "self-sacrifice": r = .41, 95%CI (.14, .62); "other-outcome": r = .35, 95%CI (.07, .58)). While Experiment 1a left some ambiguity as to whether participants were reasoning about pairwise relationships or general social dispositions, the stimuli in Experiment 1c made it impossible for participants to infer specific pairwise relationships, as only one agent was depicted. Thus, the strong quantitative fit of our main model, and the similarity of results between Experiment 1a and 1c, both suggest that participants used the same underlying utility-based reasoning when making inferences about specific pairwise relationships, as well as general social dispositions.

3.4. Experiment 2

Experiment 1 established that people can use information about an agent's personal choices and social knowledge to infer their social preferences. However, our computational framework further predicts that people should be able to perform other kinds of inferences from the same generative model of agent behavior. In Experiment 2, we tested a reverse inference to the one from Experiment 1. Here we considered cases where an agent's social preferences were known, and tested whether people could use this information to infer the agent's social knowledge based on how they behave (e.g., if A is known to care about others, but takes a seemingly inconsiderate action, what does that reveal about their knowledge?).

3.4.1. Participants

40 adult participants with US-based IP addresses were recruited via Amazon Mechanical Turk (mean age = 38.7, S.D. = 10.6). Five additional participants were recruited but excluded for failing at least one comprehension check question on each of two attempts.

3.4.2. Stimuli

Experiment 2 comprised 25 trials, each of which depicted two agents A and B, and a box containing two burgers and two veggie wraps (four items total) from which each agent could take two items. The stimuli depicted A's choice of items to keep, A's personal preference ("I prefer [burgers/veggie wraps] over [veggie wraps/burgers]", or "I like burgers and veggie wraps equally"), as well as A's social preference towards B ("A cares about him/herself more than he/she cares about B", "A cares about B more than he/she cares about him/herself", and "A cares about him/herself and B about equally"). The stimuli also depicted A's choice, but unlike the Experiment 1 stimuli, did not

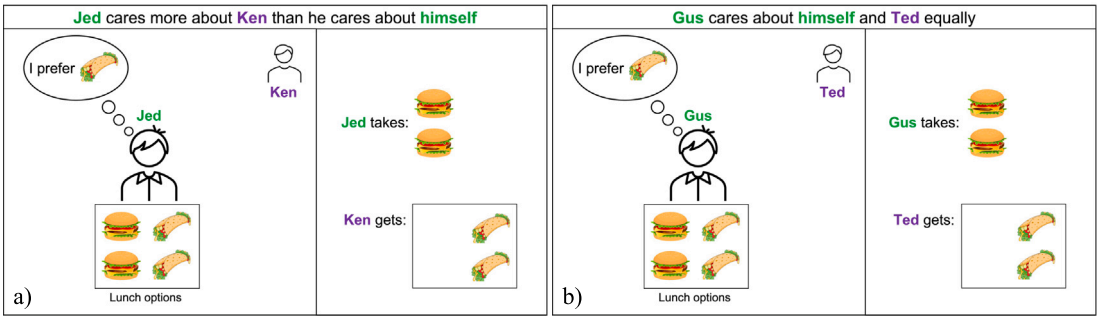

Fig. 9. Examples of stimuli from Experiment 2. Each stimulus depicts the chooser's personal preference (either "I prefer [burgers/veggie wraps]" or "I like both equally"), social preference (which agent A cares about more), and choice (which two items to keep).

provide any information regarding A's knowledge of B's preference. See Fig. 9 for examples of stimuli from Experiment 2.

Each trial varied A's personal preference (3 conditions), A's social preference towards B (3 conditions), and A's choice of food items (3 conditions: two of one item, two of the other, or one of each), yielding 27 total trials. Before collecting any data, we applied our inference model to each trial, computed the model's likelihood of A's choice in that trial, and eliminated any trial in which A's choice was highly unlikely ($< 5\%$). We took this step to avoid showing participants any scenarios that lacked a coherent social interpretation (e.g.: A is stated to be selfish but behaves in an altruistic way). This process eliminated 2 trials, leaving the 25 included in Experiment 2. (See supplemental materials for full list of trials).

3.4.3. Procedure

Participants were first shown a series of instructions providing a cover story and explaining how to interpret the stimuli. The cover story explained that A and B work in an office that provides lunch for its employees. Each morning the manager brings a box of food items (hamburgers and veggie wraps), and each employee is allowed to take two items. The story further explains that A's lunch break is earlier than B's, so that A gets first choice. After reading the instructions, participants were given an 8 question comprehension check. Participants who failed at least one question were shown the instructions a second time, followed by a second attempt at the comprehension check. Participants who failed at least one question on the second attempt were excluded from the study.

All other participants were then shown all 25 trials in a random order. After each trial, participants were asked two questions. First, participants were asked whether they thought that A knew what B preferred, and responded using a continuous slider from 0 (definitely does not know) to 1 (definitely knows). Second, participants were told to suppose that A did, in fact, know what B preferred, and asked what A thought B preferred. Participants responded using a continuous slider that depicted one food item (hamburger) on the leftmost side of the scale, the other food item (veggie wrap) on the rightmost side of the scale, and "B likes both equally" in the center of the scale. Participant responses to the second question were converted to a real number in the interval $[-1, 1]$ for analysis.

3.4.4. Model predictions and alternate models

Predictions for Experiment 2 were generated using Eq. (2.5), which takes A's reward function R_A, A's social preference W, and A's decision d, and returns a probability distribution $P(K|d, W, R_A)$ over A's beliefs K. To implement this model, we converted the verbal descriptions of A's preferences into a probability distribution over reward functions following the same method from Experiment 1a. We used a similar method to convert the verbal description of A's social preferences (e.g.: "A cares more about B than about [him/her]self") into a probability distribution over social preferences: we started with a uniform prior distribution over all possible social preferences $W = (w_A, w_B)$, then used Bayesian posterior updating to condition on the observation that $w_A > w_B$ (for "A cares more about [him/her]self"), $w_A < w_B$ (for "A cares more about B"), or $w_A = w_B$ (for "A cares about both equally") (see supplemental materials for full details).

In addition to the main model, we generated predictions using three alternate models. First, we tested a "no-discounting" model, to evaluate whether people's belief inferences also reflected an expectation of discounted rewards. Second, we wished to evaluate whether people were actually incorporating A's reward function and decision into their inferences, or simply basing their inferences on A's social preference (e.g.: assuming that A is more likely to know B's preference when A cares about B, but ignoring all other information — see 2.2). To this end, our second alternate model ("prior-only") generated predictions based solely on A's social preference, without incorporating any other information in the stimulus. Finally, to evaluate whether this non-uniform knowledge prior was, in fact, reflected by participants' judgments, we tested a "uniform-prior" model, which performs the same inference as the full model, but assumes that A is equally likely to know or not know B's preference, regardless of A's social preference towards B.

3.4.5. Results

Participant judgments and model predictions were processed in the same way as Experiments 1a-1c, as preregistered (see osf.io/56xzy). As shown in Fig. 10, main model predictions closely matched average participant judgments for both parameters (P(Knows?): $r = .95$, 95% CI(.88, .98), P(Belief): $r = .91$, 95% CI(.81, .96), combined: $r = .91$, 95% CI(.81, .96)), and were significantly more correlated with participant judgments than all three alternate models, defined as a bootstrapped difference in correlations with 95% CI not crossing 0 (see Supplemental Materials).

Participants were therefore able to leverage information about A's personal preferences, social preferences, and decisions to infer both (a) whether A has knowledge of B's preference and (b) what that knowledge is (assuming A has any) in a fashion that reflects expected utility computations. To better illustrate how our model captured participant intuitions, Fig. 11 shows the results for each trial, and highlights two pairs of trials that demonstrate two key qualitative predictions of our model. The first pair, highlighted in red (first row of Fig. 11, second and third column from the left) demonstrate how A's social preference affected people's inferences about A's social knowledge. In these two events, A took the two burgers which she preferred, and left the two veggie wraps for B. When A was known to be selfish (left trial), A's decision did not reflect their knowledge or ignorance about B's preferences at all, since A would most likely make the selfish decision regardless of B's preferences. In this case, the model relied solely on the knowledge prior (i.e.: the expectation that selfish agents are less likely to track the preferences of others), and thus inferred that a selfish A most likely did not know B's preferences, consistent with participants'

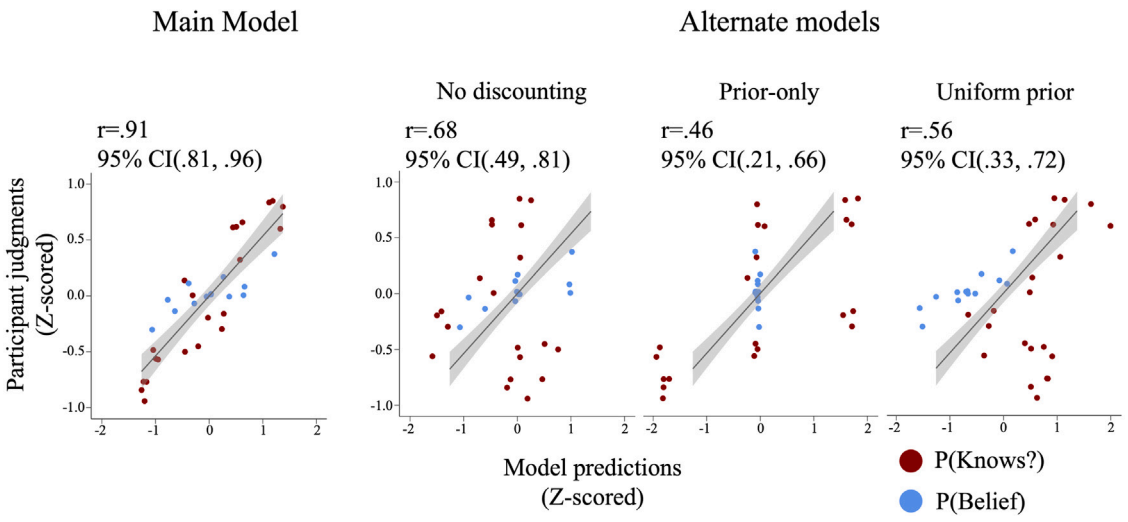

Fig. 10. Results from Experiment 2. Each subplot shows a comparison between predictions generated by one of the four models (including the three alternate models) against participant responses. Each point represents a judgment from a single trial: red dots correspond to the probability that A knows B's preference, and blue dots correspond A's belief about B's preference (assuming A *does* know). The *x*-axis depicts model predictions (z-scored) and the *y*-axis depicts participant responses (z-scored within participant and averaged across participants). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

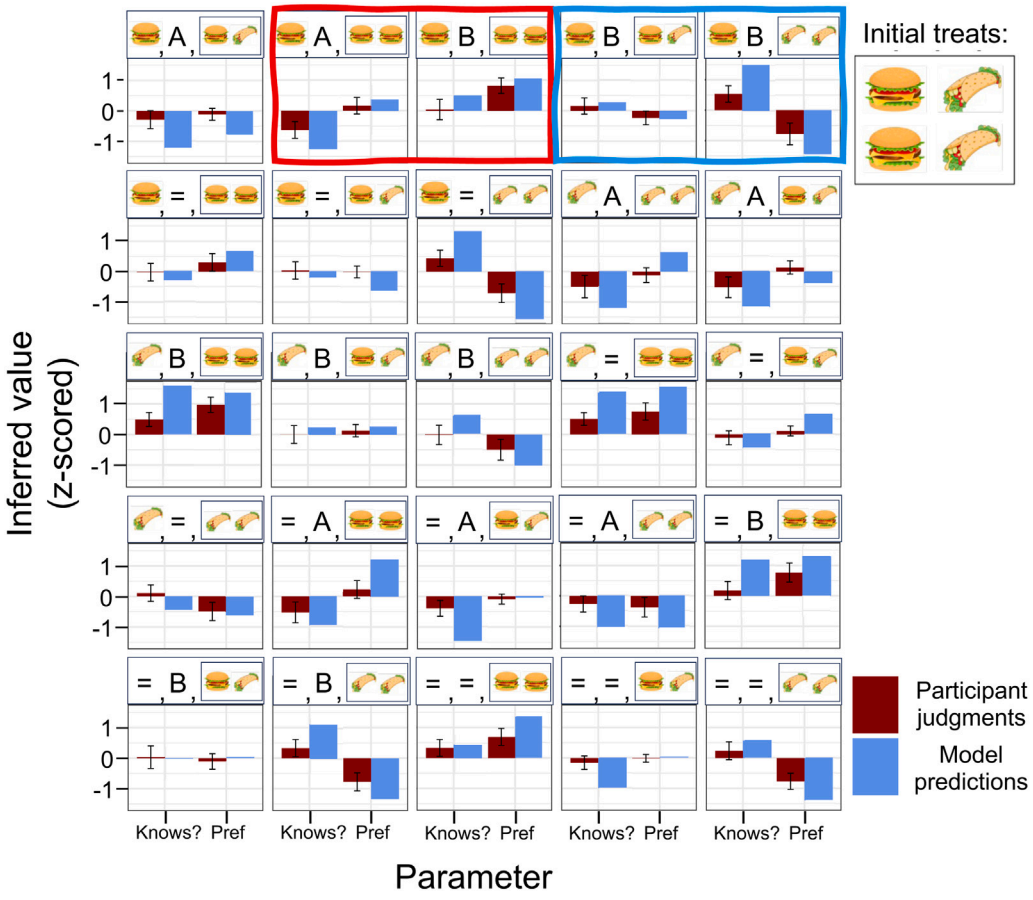

Fig. 11. Model predictions and mean participant judgments for all 25 trials in Experiment 2. Trial configurations are shown at the top of each chart. Reading left to right, the headings indicate (1) A's personal preference, (2) which agent A cares more about, and (3) which treats A chose to keep. "Knows?" is the probability that A knows B's preference, while "Pref" is B's inferred preference (assuming A does know it). Bar charts show z-scored model predictions compared against participant judgments (z-scored within participants and averaged across participants). Vertical bars show bootstrapped 95% confidence intervals. The two pairs of trials highlighted in red and blue demonstrate two key qualitative effects captured by our model, as explained in Section 3.4.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

inferences. By contrast, when A was known to be prosocial (right trial), but took an apparently selfish action (keeping both hamburgers), both participants and the model inferred a higher likelihood that A believed B preferred veggie wraps.

This result might suggest that participants simply inferred that a prosocial A is generally more likely to know B's preference, but the next two examples, highlighted in blue (first row of Fig. 11, fourth and fifth column from the left), showed this was not the case. When A is shown to be prosocial, and took the "choice-maximizing" action, (i.e.: taking one of each — left trial), both the model and participants expressed uncertainty about A's knowledge. Even though A was prosocial, which is associated with a higher overall likelihood of knowing B's preference, A's action suggested that A did not, in fact, know B's preference, resulting in greater uncertainty about A's social knowledge. However, when A took both of their less preferred snacks (veggie wraps) and left their more preferred snacks (hamburgers) for B (right trial), participants and the model both inferred that A knew what B preferred. Thus, participants leveraged an expectation of prosocial behavior to infer A's knowledge or ignorance of B's preference, as predicted by the model. This demonstrates that participants could not only infer social preferences from actions and social knowledge, but also infer social knowledge from actions and social preferences.

The results of our three alternate models also emphasized how each of the core assumptions in our main model was necessary to account for participant intuitions. First, the "no-discounting" model (r = .68, 95% CI(.49, .81)) failed to predict participant inferences about A's knowledge or ignorance of B's preferences (i.e. P(Knows?). This is because, when A did not know B's preferences, all three possible decisions yielded the same expected reward to B in the absence of discounting, so the "no-discounting" model could not infer whether or not A knows B's preference (see results section of Experiment 1a for an extended explanation of this point). Second, the significantly lower correlation of the "prior-only" model (r = .46, 95%CI (.21, .66)) suggested that people were not just assuming that prosocial agents are more likely to know the preferences of others, but that they also incorporated information about agents' rewards and decisions into their inferences. Finally, the "uniform prior" model, which assumed that A's social preference has no effect on the likelihood that A knows B's preference, had a significantly lower correlation with participant data than the main model (r = .56, 95%CI (.33, .72)). This suggests that participants did expect that people with prosocial preferences were more likely to know the preferences of others. Thus, these results strongly suggest that (a) participants could infer an agent's social knowledge based on their social preferences and decisions, (b) that these inferences reflected the same discounted utility computations underlying the social preference inferences from Experiment 1, and (c) participants assumed that prosocial agents were more likely to know the preferences of others, but still attended to the costs they incurred and rewards they yielded when making these social judgments.

4. Discussion

How do we detect potential social partners who are kind, empathetic, and supportive in everyday social life? Here we examined the idea that people can detect these traits from indirect evidence of prosociality revealed by agents' personal choices. We presented a computational model of this capacity, focusing on settings where agents' personal choices can signal that they deployed theory of mind to indirectly benefit others. At the heart of this model is an assumption that people expect agents to maximize their own utilities, including the utility they assign to others' welfare.

Across four experiments, we found converging support for our computational account. In Experiments 1a-1c, we found that people leverage information about an agent's personal preferences, choices, and knowledge of others' preferences to infer that agent's social preferences. We found that people can make these same evaluations using both direct, qualitative evidence of preferences (Experiment 1a) as well as indirect, quantitative evidence of preferences from choice histories (Experiment 1b), and that people make very similar judgments in these cases for both specific pairwise relationships (e.g.: how much does A care about B; Experiments 1a & 1b) as well as general social dispositions (e.g.: how much does A care about others; Experiment 1c). In Experiment 2, we found that the inverse is also true: people can infer an agent's knowledge (or ignorance) of others' preferences based on the agent's social preferences and personal choices, and these inferences reflect the same underlying utility computations as our initial experiments. Together, this work supports the idea that people can draw on indirect evidence of prosociality (making personal choices that indirectly fulfill others' preferences) to discern an agent's social preferences. This provides strong support for a mentalistic, utility-based account of social inference, and emphasizes the key role of detecting when an agent is using theory of mind in the service of others.

A second, somewhat more surprising finding of our experiments was the role of discounting in producing the intuitive "social mindfulness" effect in these tasks. In particular, previous research suggests that, in the absence of any information about B's preferences, there is a strong intuition that the prosocial decision is the one that maximizes the number of available options for B (i.e.: for A to leave one of each type of item for B) (Van Doesum et al., 2013; Zhao et al., 2021). This effect was also found across both of our studies, despite the fact that B did not technically have a choice to make in our task (as B automatically got everything that A left behind). However, our analysis revealed that this social mindfulness effect was not predicted by an alternate model in which reward functions were not discounted. That is, when A was totally ignorant of B's preference, then the expected reward to B was the same regardless of A's choice, so the "no-discounting" model could not distinguish A's social preference in these cases. Only when the individual reward functions were discounted did the social mindfulness effect emerge. This finding highlights the value of computational models in social psychology: while the social mindfulness intuition is fairly salient and immediate, it appears to be driven in these cases by a separate intuition regarding diminishing marginal utilities of consumable goods. This also points to potential future studies to further explore the role of discounting in social judgments. If, for example, we translate the task to another domain with no diminishing marginal utilities (e.g.: if I like one genre of book over another, I might enjoy a second book from the same genre just as much as the first one), we might expect qualitatively different intuitions about prosocial behavior in that context.

While these results are promising, they also have some limitations. First, our work focused on tasks in which choices were concrete and involved only a narrow set of options. In real-world scenarios, however, people often encounter 'tests' of social mindfulness where the choice options themselves are not clearly visible, but must be represented abstractly at the time of choice. For instance, choosing to keep the seat next to one clear on a crowded bus, choosing to keep one's desk clean in a shared office space, or choosing to shop at a local farmer's market as opposed to a large chain supermarket. Generalizing our framework to a wider range of more naturalistic task settings is therefore an important step for future research.

Another limitation is that our work did not conclusively determine how people think about the relationship between prosociality and knowledge. In particular, our model assumed that more prosocial agents are more likely to attend to and know others' preferences. Our model's high quantitative fits, and the lower fits when this prior is removed, support this idea (see Experiment 2). However, we do not know whether people also expect agents with malicious or manipulative intents to also be more likely to attend to and know others' preferences. This is because cases where agents are antagonistic fell outside the range of our experiments and model. Determining whether the expectation of knowledge also extends to agents with malicious attitudes would require a paradigm where agents are able to manipulate each other. This is a direction that we hope to explore in future work.

Additionally, the present work considered individual instances of choice outside of a broader social context. In analogous real life settings, numerous factors may color a person's social inferences. One important factor not considered in the present work is how reputational

concerns might affect such social judgments. In many situations, decisions that benefit others (even in subtle, indirect ways) are driven not just by altruistic motives, but also from a desire to improve (or maintain) a positive reputation with others. Indeed, existing work suggests that social behavior is highly influenced by reputational concerns (Ariely, Bracha, & Meier, 2009; Asaba & Gweon, 2022), and that people are sensitive to such influences when making inferences about why agents engage in prosocial behavior (Barasch, Levine, Berman, & Small, 2014; Carlson & Zaki, 2018). Thus, an important extension of our model would be to account for how observers consider reputational factors or other kinds of social pressures that might shape an agent's behavior.

Another crucial factor is the perceiver's own model of the social environment in which the interactions take place. For example, group identities and intergroup dynamics can powerfully shape people's perceptions and expectations about how agents should (or do) treat each other (Lees & Cikara, 2020; Rhodes, 2013; Waytz, Young, & Ginges, 2014). Moreover, social inferences surrounding a helpful act might be colored by very different expectations if an observer believes an actor and recipient belong to the same ingroup, versus if one is an ingroup member and one is an outgroup member. Accounting for the observer's knowledge about the relevant social environment is another important next step for future research. Conversely, it may also be possible to perform the reverse inference — that is, to reason about the underlying structure of social groups by observing which agents do or do not act prosocially towards one another. Suppose, for example, an observer believes that ingroup members are more likely to behave prosocially towards each other than towards outgroup members. The observer could apply a model like the one proposed here to infer dyadic social preferences between pairs of individuals, then extrapolate the broader social structures from these dyadic relations (Davis, Dunham, & Jara-Ettinger, 2022; Gershman & Cikara, 2020).

How do people detect potential social partners? Here we proposed that part of the solution is a sensitivity to how agents make personal decisions in situations where their choices might indirectly benefit others. Across a series of experiments, we showed that people not only leverage information about an agent's personal choices to infer their social preferences, but also leverage information about an agent's social preferences to infer their knowledge about the preferences of others. These findings showcase the promise of formal approaches for a deeper understanding of social psychology. Our future work will further explore the interactions between inferences about dyadic social relations and broader social structures, as well as the role of reputational concerns in such inferences. Indeed, by more closely examining how factors such as reputation, group dynamics, and choice representations are integrated into people's social inferences, future work will yield richer accounts of how people discern others' social preferences, and in turn, how people decide who to trust and befriend in social life.

CRediT authorship contribution statement

Isaac Davis: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Ryan Carlson:** Conceptualization, Methodology, Writing – original draft. **Yarrow Dunham:** Conceptualization, Methodology, Writing – review and editing, Funding acquisition. **Julian Jara-Ettinger:** Conceptualization, Methodology, Writing – review and editing, Funding Acquisition.

Data availability

Data will be made available on request.

Acknowledgments

We are grateful to the editor and three anonymous reviewers for their helpful feedback on this manuscript, as well as the Computational Social Cognition and Social Cognitive Development labs for their continued input and support. This work was supported by National Science Foundation (NSF) awards IIS-2106690 and BCS-2045778.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105580.

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), 544–555.

Asaba, M., & Gweon, H. (2022). Young children infer and manage what others think about them. *Proceedings of the National Academy of Sciences*, *119*(32), Article e2105642119.

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 the 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.

Berman, J. Z., & Silver, I. (2022). Prosocial behavior and reputation: When does doing good lead to looking good? *Current Opinion in Psychology*, *43*, 102–107.

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. (2022). How inferred motives shape moral judgements. *Nature Reviews Psychology*, 1–11.

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

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320–17325.

Davis, I., Dunham, Y., & Jara-Ettinger, J. (2022). Inferring the internal structure of social collectives. In *Proceedings of the annual meeting of the cognitive science society, Vol. 44*.

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

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.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

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

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

Gergely, G., & Csibra, G. (2003b). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gershman, S. J., & Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science*, *29*(5), 460–466.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*(1), 30–39.

Hartman, R., Blakey, W., & Gray, K. (2022). Deconstructing moral character judgments. *Current Opinion in Psychology*, *43*, 205–212.

Heyman, G., Barner, D., Heumann, J., & Schenck, L. (2014). Children's sensitivity to ulterior motives when evaluating prosocial behavior. *Cognitive Science*, *38*(4), 683–700.

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*, Article 101334.

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.

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.

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, *4*(3), 279–286.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., et al. (2014). The child as econometrician: A rational model of preference understanding in children. *PLoS One*, *9*(3), Article e92160.

Margoni, F., & Surian, L. (2022). Judging accidental harm: Due care and foreseeability of side effects. *Current Psychology*, *41*(12), 8774–8783.

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

Morelli, S. A., Leong, Y. C., Carlson, R. W., Kullar, M., & Zaki, J. (2018). Neural detection of socially valued community members. *Proceedings of the National Academy of Sciences*, *115*(32), 8149–8154.

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*.

Nobes, G., & Martin, J. W. (2022). They should have known better: The roles of negligence and outcome in moral judgements of accidental actions. *British Journal of Psychology*, *113*(2), 370–395.

Poorthuis, A., Thomaes, S., Denissen, J., van Aken, M., Orobio de Castro, B., et al. (2012). Prosocial tendencies predict friendship quality, but not for popular children. *Journal of the Experimental Child Psychology*, *112*(4), 378–388.

Powell, L. J. (2021). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*, Article 17456916211048487.

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

Siem, B., & Stürmer, S. (2018). Attribution of egoistic versus altruistic motives to acts of helping. *Social Psychology*.

Son, D., & Padilla-Walker, L. M. (2020). Happy helpers: A multidimensional and mixed-method approach to prosocial behavior and its effects on friendship quality, mental health, and well-being during adolescence. *Journal of the Happiness Studies*, *21*(5), 1705–1723.

Swap, W. C. (1991). When prosocial behavior becomes altruistic: An attributional analysis. *Current Psychology*, *10*(1), 49–64.

Tauber, S., & Steyvers, M. (2011). Using inverse planning and theory of mind for social goal inference. In *Proceedings of the annual meeting of the cognitive science society, Vol. 33*.

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.

Van Doesum, N. J., de Vries, R. E., Blokland, A. A., Hill, J. M., Kuhlman, D. M., Stivers, A. W., et al. (2020). Social mindfulness: Prosocial the active way. *The Journal of Positive Psychology*, *15*(2), 183–193.

Van Lange, P. A. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of the Personality and Social Psychology*, *77*(2), 337.

Van Lange, P. A., & Van Doesum, N. J. (2015). Social mindfulness and social hostility. *Current Opinion in Behavioral Sciences*, *3*, 18–24.

Waytz, A., Young, L. L., & Ginges, J. (2014). Motive attribution asymmetry for love vs. hate drives intractable conflict. *Proceedings of the National Academy of Sciences*, *111*(44), 15687–15692.

Wellman, H. M. (2014). *Making minds: how theory of mind develops*. Oxford University Press.

Zang, L., Li, D., & Zhao, X. (2023). Preference matters: Knowledge of beneficiary's preference influences children's evaluations of the act of leaving a choice for others. *Journal of the Experimental Child Psychology*, *228*, Article 105605.

Zhao, X., Zhao, X., Gweon, H., & Kushnir, T. (2021). Leaving a choice for others: Children's evaluations of considerate, socially-mindful actions. *Child Development*, *92*(4), 1238–1253.