# Rational inferences about social valuation

# Rational inferences about social valuation

**Tadeg Quillien**

Department of Psychological & Brain Sciences

**John Tooby**

Department of Anthropology

**Leda Cosmides**

Department of Psychological & Brain Sciences

University of California Santa Barbara

## Author Note

This is the final submitted version of a paper published in *Cognition*. We have no conflicts of interest to report. Correspondence concerning this article should be addressed to Tadeg Quillien, current affiliation: School of Informatics, University of Edinburgh.
E-mail: tadeg.quillien@gmail.com.

**Abstract**

The decisions made by other people can contain information about the value they assign to our welfare—for example how much they are willing to sacrifice to make us better off. An emerging body of research suggests that we extract and use this information, responding more favorably to those who sacrifice more even if they provide us with less. The magnitude of their trade-offs governs our social responses to them——including partner choice, giving, and anger. This implies that people have well-designed cognitive mechanisms for estimating the weight someone else assigns to their welfare, even when the amounts at stake vary and the information is noisy or sparse. We tested this hypothesis in two studies (N=200; US samples) by asking participants to observe a partner make two trade-offs, and then predict the partner's decisions in other trials. Their predictions were compared to those of a model that uses statistically optimal procedures, operationalized as a Bayesian ideal observer. As predicted, (i) the estimates people made from sparse evidence matched those of the ideal observer, and (ii) lower welfare trade-offs elicited more anger from participants, even when their total payoffs were held constant. These results support the view that people efficiently update their representations of how much others value them. They also provide the most direct test to date of a key assumption of the recalibrational theory of anger: that anger is triggered by cues of low valuation, not by the infliction of costs.

*Keywords:* Social cognition; computational modeling; emotion; welfare trade-offs; evolutionary psychology

## Rational inferences about social valuation

### Introduction

Organisms in a social species often face trade-offs between their own welfare and that of other individuals. That is, your actions may generate costs and benefits for other individuals in addition to costs and benefits for you. How do individuals arbitrate such trade-offs? Evolutionary biologists have studied this question extensively with the help of simple game-theoretic models. They have uncovered many selection pressures that shape the behavioral strategies of agents facing welfare trade-offs (Maynard-Smith & Price, 1973; Trivers, 1971; Axelrod & Hamilton, 1981; Nowak, 2006; Hamilton, 1964; Debove et al., 2015; Quillien, 2020; Hammerstein & Parker, 1982). For example, opportunities for repeated interactions can favor the evolution of reciprocity-based strategies such as 'Tit-for-Tat': help someone if they also helped you (Trivers, 1971; Axelrod & Hamilton, 1981; Schmid et al., 2021).

When applying the insights of this research to human social behavior, the question arises of how a given behavioral strategy is implemented at the cognitive level (Cosmides & Tooby, 1992; Cosmides et al., 2010; Lieberman et al., 2007). For example, what counts as 'reciprocity' to the human mind? Human reciprocity is more complex than a simple 'Tit-for-tat' rule: when we decide whether to help someone, we care not just about what they did, but also whether their past actions reveal that they are *disposed* to help us (Delton et al., 2012; Lim, 2012; Hoffman et al., 2015).

A concern with why other people do what they do is a pervasive feature of human social cognition, across domains such as partner choice, reciprocity, conflict and moral judgment (Hoffman et al., 2015; Tooby et al., 2008; Uhlmann et al., 2015; Carlson et al., 2022; Jordan et al., 2016; Eisenbruch & Krasnow, 2022; Woo & Spelke, 2023). In particular, we care about whether someone else is disposed to incur costs to benefit us, and to refrain from hurting us for their own benefit. More formally, we represent the weight that this person assigns to our welfare when they make decisions (Lim, 2012; Tooby et al.,

2008; Eisenbruch & Krasnow, 2022; Sznycer et al., 2019; Sell et al., 2017).

For instance, if someone grabbed your scarf to wipe tomato juice off her face, you would feel angry. But if she instead used your scarf to make a tourniquet that stops her child's wound from bleeding, you would feel less anger—or none at all (Sell et al., 2017). Even though the two actions inflict the same cost on you (your scarf is stained), they reveal very different things about how much this person values you. By using your scarf as a napkin, she is imposing that cost on you to get a trivial benefit for herself. You can infer that she is likely to treat you badly in future situations as well. But if ruining your scarf was her only option to save her child's life, you might not expect her to treat you badly when she has much less at stake.

In this paper, we report evidence that people infer the weight someone else assigns to their welfare in a statistically rational manner, even on the basis of sparse data. These inferences also appear to be an input to the emotion of anger.

**The psychology of welfare-tradeoffs**

The way humans make welfare-tradeoffs depends on who they are interacting with. For instance, you may offer to take a friend to the airport at 6am, but might not make the same offer to an acquaintance. To a good approximation, these decisions can be modeled by assuming that the decision-maker puts some weight on the payoffs of the other individual (Delton et al., 2023; for review of the empirical evidence see Delton and Robertson, 2016). In this view, you take your friend to the airport at a cost to yourself because you value the fact that this increases their welfare; you might not help an acquaintance in the same situation, because the weight you put on their welfare relative to your own is lower than the weight you put on your friend's welfare (Eisenbruch & Krasnow, 2022; Delton, 2010).

Switching perspectives, other individuals vary in how much weight they put on your welfare: Your friends may sacrifice more to help you than acquaintances will, strangers may not care about you, and your rivals might actively try to hurt you. It pays to know

who values you and by how much, in order to know who to avoid, who to associate with, and when you should try to bargain for better treatment (Sell et al., 2017; Sell et al., 2009; Barclay, 2013; Baumard et al., 2013).

By contrast, evaluating other individuals on the basis of a simple tally of the costs and benefits they generate for you (as in model-free reinforcement learning, Sutton and Barto, 2018) is not optimal (Qi & Vul, 2022). The payoffs you get in an interaction with someone are not necessarily good predictors of future payoffs from interacting with that individual: For example, someone who failed to help you when they were ill might still help you in the future when they have recovered.

If this perspective is correct, evolution should have designed cognitive mechanisms in humans that can infer how much other individuals value your welfare (Lim, 2012; Tooby et al., 2008; Sell et al., 2017). Formal evolutionary models support this argument (Qi & Vul, 2022) and suggest that inferences about social valuation should play a large role in how people evaluate others, even relative to other factors such as perceived competence (Eisenbruch & Krasnow, 2022).

Indeed, many aspects of human social cognition appear to be regulated by inferences about how much someone values your welfare. When people have to choose between partners, they prefer those who appear to value them—even over potential partners who generate more resources for them but value them less (Lim, 2012; Eisenbruch and Roney, 2017; Hackel et al., 2015; Hackel et al., 2020; Raihani and Barclay, 2016, see also Sznycer et al., 2019; Dhaliwal et al., 2022). Social perception tends to prioritize traits such as “warmth” or “generosity” (which in many cases track how much someone values your welfare) over traits related to competence (reviewed in Eisenbruch & Krasnow, 2022; Fiske et al., 2007). People are interested in learning about events that can tell them how much someone else is willing to help (Quillien, 2023) and automatically categorize others on this basis (Delton & Robertson, 2012). Representations of welfare valuation might already be present in infancy (Powell, 2021).

In turn, inferences about social valuation appear to be an input to social emotions like anger and gratitude (Sznycer et al., 2021; Monroe, 2020). Many features of anger are well-explained by the hypothesis that anger is a neurocomputational system that evolved to "bargain" for better treatment when another person seems to be putting too little weight on your welfare (Sell et al., 2017; Sell et al., 2009). Similarly, gratitude may serve to consolidate valuable social relationships, by signaling that you recognize that the other person values you highly (Lim, 2012; Algoe, 2012; Smith et al., 2017).

This body of work raises an important question: How do people infer how much weight someone puts on their welfare? Similar to vision and other basic perceptual processes (Knill & Richards, 1996; Helmholtz, 1856), social perception takes place in a highly uncertain world, and must rely to a great extent on statistical inferences. For instance, the observation that Alice did not share her cake with you is consistent with many hypotheses about how much she values you: Maybe she wants you to stay hungry; maybe she likes you but cares about herself more; maybe she mistakenly thought you were on a diet. This one decision by Alice is not enough to infer the exact value of the weight that she puts on your welfare, but it can be used to narrow down a probabilistic estimate of that weight.

**The present research**

Studying the inference problem we just described allows for a stronger test of the existence of representations of social valuation in humans. We developed a task in which participants need to predict the behavior of other players in a simple economic game. Participants were paired with partners who could decide to allocate money either to themselves or to the participant. After first briefly observing one of their partners play two rounds of this game, participants were asked to predict what that same person would do in other rounds of the game. As an example (see Figure 1), suppose that you are told that in a given round of the game, your partner had a choice between getting $12 for herself or

letting you have $31, and that she decided to let you have $31. In another round of the game, where she had a choice between $27 for herself or $30 for you, how likely is it that she decided to allocate the money to you?

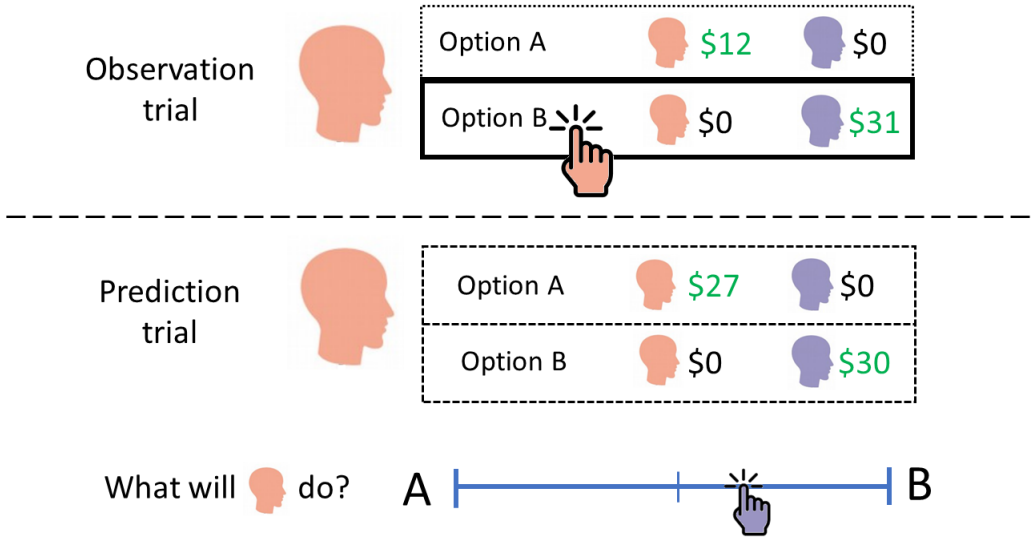


**Figure 1**

***Schematic description of our prediction task.*** *The participant is in blue, their partner is in red. In observation trials, the participant observes the partner make a decision. In prediction trials, the participant reports their probabilistic belief about what their partner will do. There were two observation trials and five prediction trials for each partner (not shown).*

This is a difficult problem of prediction under uncertainty, because participants must generalize on the basis of very sparse data. To generate predictions, they need to rely on inductive biases, in the form of a causal model of how people typically make decisions in situations involving welfare-tradeoffs. Our hypothesis is that this causal model relies on representations of the weight that the decision-maker assigns to the other person's welfare.

To test this hypothesis, we compared participants' predictions to those made by a Bayesian ideal observer model, which makes predictions on the basis of representations of social valuation. An ideal observer model makes predictions in the statistically optimal

way (Geisler, 2011), given some assumptions about how the world works (here, assumptions about how agents typically make welfare trade-offs). Importantly, the presence of these assumptions in our ideal observer correspond to substantial hypotheses about the causal model that people bring to the prediction task.

Formally speaking, we assume that people have an implicit causal theory of other agents as utility-maximizers who (potentially) attach some weight to the utility of a given interaction partner. People use this causal model to predict how someone will behave in situations that involve potential costs and benefits for other parties. They can also use this causal theory to learn: after observing an individual A do something helpful or selfish toward B, they update their estimate of the weight that A assigns to B's welfare (see methods for details).

We also make the critical assumption that people make inferences that are approximately rational. Given the evolutionary importance of inferences about valuation, the mind should have cognitive machinery that makes these inferences efficiently, in a way that approximates the normative standards of probability theory. Note that this prediction is in stark contrast to a decades-long research program in cognitive psychology that has documented multitudes of contexts in which humans depart from Bayesian rationality (Kahneman et al., 1982; Marcus, 2009).

In existing studies using ideal observer models of social perception (e.g. Qi & Vul, 2022; Xiang et al., 2013; Siegel et al., 2018; Barnby et al., 2020), people are typically given very rich information during the learning process. But many different learning algorithms can perform well given enough information. The fingerprints of Bayesian updating are clearest when an agent needs to make inferences from sparse data. So we asked participants to make predictions about a partner after seeing only two choices that this partner had made. This allows for a strong test of our hypothesis: participants should make inferences efficiently, based even on thin slices of behavior.

That is, we used the ideal observer's predictions as a benchmark, to assess how well

people infer how much someone values them, based on sparse data. This mirrors real-life: It can be important to estimate how much someone values you even early in a relationship, when opportunities to observe them making welfare tradeoff decisions toward you are limited.

The ideal observer framework also allows a strong test of the hypothesis that anger (/gratitude) is elicited by actions that suggest that the actor puts a low (/high) weight on your welfare. We test this proposal in a more direct way than existing studies (e.g. Lim, 2012; Sell et al., 2017), by looking at whether the inferences drawn by the ideal observer predict how much anger or gratitude participants report in response to a given decision, above and beyond the payoff consequences of that decision.

## Method

### Overview of the task

Participants played a simple economic game, the Welfare-Tradeoff Task (WTT), with a series of partners. The WTT is a dictator game with binary choices (Delton, 2010). In a trial of the WTT, if Alice is the dictator and Bob is the recipient, Alice must choose between two alternatives, where $\pi$ is a payoff in dollars:

- Alice receives $\pi_{\text{alice}}$, Bob receives \$0
- Alice receives \$0, Bob receives $\pi_{\text{bob}}$

Participants played the WTT in the role of the receiver. For each partner they were paired with, they saw the decisions that their partner had made in two rounds of the game (see Figure 1, top for a schematic illustration of one of these *observation trials*).

After these two 'observation' trials, participants saw the options that had been presented to the same partner in 5 other WTT trials. For example, they might see a trial in which their partner had a choice between \$27 for herself and \$30 for the participant (see bottom of Figure 1). For each of these trials, participants were asked to predict the

probability the partner chose to allocate money to the participant, using a slider scale from 0% to 100% likely.

Additionally, we asked participants how angry and how grateful they felt toward their partner, after having seen what they did in the two observation trials (but before the prediction trials).

**Procedure**

After participants signed a consent form, they read a description of how the WTT works[1]. To familiarize themselves with the WTT, participants played four rounds of a pretend version of the game in the role of dictator, while being asked to imagine that the receiver was one of their acquaintances. Throughout the study, no money was involved, but participants were asked to imagine that they were playing for real money.

In the main task, participants played the WTT in the role of the receiver. The dictators they played with were sham partners generated by the computer. Participants were aware of this; no deception was involved at any point in the study. We asked participants to imagine that these partners were acquaintances, each a different one.

Each participant played the WTT with 10 partners in total. Partners were always the dictator, while participants were receivers. Participants were referred to as "you". For each partner, you first see the decisions the partner made on two WTT trials. For example, you might see that your partner had a choice between \$12 for himself and \$31 for you, and decided to allocate \$31 to you. You then see a second choice made by the same partner; e.g., when choosing between \$2 for himself and \$92 for you, this partner decided to allocate \$92 to you. After you have observed the two decisions, you are asked how grateful and how angry you feel toward the partner, using 1-7 likert scales.

After these two 'observation' trials, you are shown the choices presented to that

[1] The description specified that when two people play several rounds of the WTT, only one round will be randomly selected to be paid out, so that players should treat each trial as if it was the only one.

partner in 5 other WTT trials. For example, you may see a trial in which your partner had a choice between $27 for himself and $30 for you. For each of these trials, you are asked to predict the probability the partner chose to allocate money to you, using a slider scale from 0% to 100% likely. We counterbalanced the framing of the question such that half the participants were actually asked to rate the probability that the partner would allocate the money to themselves (i.e. the partner), and reverse-coded the ratings for these participants. After completing these five prediction questions, the participant observed and made predictions about a different partner.

Each trial was displayed on a separate page; each page also displayed, as a reminder, the two decisions that the participants had initially observed. We did not give feedback about the accuracy of the participants' predictions.

Partners were presented in random order. Among the 10 partners, 5 were "generous" in the two observation trials (they allocated money to the participant both times), and 5 were "selfish" (they allocated the money to themselves both times). Table 1 shows the choices each partner faced in the two observation trials, and whether that partner chose to give money to the participant or take money for himself. The order in which the two decisions by a given partner were presented was counterbalanced across participants.

Table 2 shows the potential payoffs for the five prediction trials. These 5 trials were identical for all partners, but the order in which they were presented was randomized (i.e. presentation order was not necessarily the same for each partner).

Finally, participants completed a few demographic questions, and were thanked for their participation.

### Materials

We designed the generous partners such that the *sum* of their two decisions had the same material consequences: As Table 1 shows, each generous partner incurred an opportunity cost of $14 (the amount forgone by allocating to the participant both times)

and delivered $125 to the participant.[2] Although each generous partner delivered the same total payoff to the participant, however, they made decisions that revealed different things about the weight they assign to the participant's welfare. Partner A, for example, is willing to forgo $29 to let the participant have $29; by contrast none of the decisions made by partner E necessarily implies a high willingness to sacrifice for the participant (see Table 1). As such, even though in total partners A and E both sacrificed $14 in order to let the participant have $125, our ideal observer model makes very different inferences about how much they value the participant.

Similarly, each selfish partner gained a total of $60 instead of delivering $35 to the participant, but they made decisions that had different implications about how much they are willing to benefit at the expense of the participant. For example in one round partner J gets $1 instead of letting the participant have $35, revealing that J cares very little about the participant. In contrast partner F, who also decides to take the money in both rounds, would have forgone large amounts of money by deciding to give.

The observation trials were constructed so that one decision conveyed little information about how much the partner values you (e.g.,they forgo $2 to let you have $92), whereas the other decision conveyed some information about how much the partner was willing to sacrifice to give you approximately $30 (e.g., they forgo $12 to give you $31).

[2] Due to an addition error when designing the study, partner D gave a total benefit of $123 instead of $125 to the participant.

| Partner | $\pi_{\text{partner}}$ | $\pi_{\text{participant}}$ | decision | ideal-observer-inferred WTR |
|---|---|---|---|---|
| A | 29 | 29 | Give | |
| | -15 | 96 | Give | 1.53 |
| B | 24 | 33 | Give | |
| | -10 | 92 | Give | 1.35 |
| C | 15 | 27 | Give | |
| | -1 | 98 | Give | 1.24 |
| D | 12 | 31 | Give | |
| | 2 | 92 | Give | 1.14 |
| E | 5 | 35 | Give | |
| | 9 | 90 | Give | 1.04 |
| F | 50 | 29 | Take | |
| | 10 | 6 | Take | 0.36 |
| G | 26 | 33 | Take | |
| | 34 | 2 | Take | 0.01 |
| H | 16 | 27 | Take | |
| | 44 | 8 | Take | -0.1 |
| I | 12 | 31 | Take | |
| | 48 | 4 | Take | -0.24 |
| J | 1 | 35 | Take | |
| | 59 | 0 | Take | -0.49 |

**Table 1**

*Observation trials. Choices each partner faced and their decisions, along with the WTR that the ideal observer model inferred based on these two decisions. The order in which the partner's decisions were presented was counterbalanced across participants.*

| $\pi_{\text{partner}}$ | $\pi_{\text{participant}}$ | $\phi$ |
|---|---|---|
| 39 | 30 | 1.3 |
| 27 | 30 | .9 |
| 16.5 | 30 | .55 |
| 7.5 | 30 | .25 |
| 1.5 | 30 | .05 |

**Table 2**

*Prediction trials. Potential payoffs for the partner and the participant, for the five prediction trials.* $\phi = \frac{\pi partner}{\pi participant}$

### Ideal observer model

The first component of the ideal observer model is a generative model of behavior in the WTT. This model can be seen as a set of assumptions about how humans typically behave in the game (for empirical evidence in favor of the model see Delton et al., 2023; Delton, 2010; for existing computational models of social cognition making similar assumptions see e.g. Jern and Kemp, 2014; Ullman et al., 2009; Davis et al., 2021). It holds that people are more likely to Give when the cost of doing so is low, but that people are more or less likely to give depending on the weight they assign to the recipient's welfare. This weight is called a *Welfare-Tradeoff Ratio* (WTR).

Formally, the generative model holds that Alice plays the WTT so as to maximize her expected utility[3], given by:

[3] Note that the WTT has a simple enough structure that the single-parameter WTR utility function is an adequate model of human behavior (see Delton, 2010), but a full cognitive model of welfare-tradeoffs in humans would require more parameters to explain behavior in more complex settings (e.g., to capture the extent to which people are sensitive to variation in the cost-effectiveness of a helpful action; see Andreoni and Miller, 2002; Fisman et al., 2007).

$$U_{\text{alice}} = \pi_{\text{alice}} + \text{WTR}_{\text{alice}\rightarrow\text{bob}} * \pi_{\text{bob}}$$

under the constraint that Alice has a noisy representation of the payoffs involved in a given trial. Specifically, for each trial she observes a value of $\phi = \pi_{\text{alice}}/\pi_{\text{bob}}$ drawn from a normal distribution with mean $\phi$ and variance $\sigma_\phi^2$. This constraint makes her choices non-deterministic, and models the fact that humans are not always perfectly consistent in their behavior when they make welfare-tradeoffs (Delton, 2010; Fisman et al., 2007)[4].

Second, the ideal observer model uses Bayesian inference to update its belief about an agent's WTR, on the basis of observations of the agent's decisions.

The ideal observer has a prior distribution over $\text{WTR}_{\text{alice}\rightarrow\text{bob}}$, which represents its initial belief about Alice's WTR, before it has had access to any specific information about Alice. When observing a decision made by Alice (give or take), the model uses Bayes' rule to 'invert' the generative model and update its belief about Alice's WTR:

$$P(\text{WTR}|\text{decision}, \phi) = \frac{P(\text{decision}|\text{WTR}, \phi)P(\text{WTR})}{P(\text{decision}|\phi)}$$

where $\phi = \pi_{\text{alice}}/\pi_{\text{bob}}$ in that trial. Concretely, the ideal observer uses its generative model in order to simulate what Alice would do if she had a given WTR toward Bob, and then compares the outcome of this simulation to what Alice actually did. The ideal observer uses this comparison to adjust how likely it is that Alice does in fact have this WTR toward Bob. It does so for a wide range of different possible WTRs (see SOM for mathematical details).

Third, the ideal observer can predict what Alice would do in a given trial of the WTT, given its belief in Alice's WTR, using the following equation:

[4] Alternatively, one could assume that Alice's choices are the output of a softmax function. We prefer the current implementation because it yields a noise parameter with a more natural interpretation, compared to the temperature parameter of a softmax. See also discussion in the Appendix of Qi and Vul (2022).

$$P(\text{decision}|\phi) = \int P(\text{decision}|\text{WTR}, \phi) P(\text{WTR}|*)\, d\text{WTR}$$

Where $p(\text{WTR}|*)$ denotes the model's posterior belief about Alice's WTR (see SOM for details).

Although the ideal observer's belief about a partner's WTR is a probability distribution, it is often more convenient and intuitive to consider it as a single number. In our analyses, when we refer to the WTR that the observer infers a partner to have, we are using the median of this distribution.

**Parameterization of the ideal observer**

The ideal observer must be equipped with a prior: a baseline expectation about the WTR of a partner for which the observer has no information. We set this prior on the basis of empirical data on participants' prior beliefs about the distribution of WTRs among their acquaintances. Specifically, we asked participants in Study 2 to complete an additional task. After completing the WTT familiarization phase, but before the prediction task, they were asked to complete a variant of the prediction task where they had to predict the behavior of 20 different interaction partners for whom they had not observed any prior decision. They made one prediction per partner, in trials of the WTT with $\pi_{\text{participant}} = \$30$ and $\pi_{\text{partner}}$ ranging from \$3 to \$60 in \$3 increments (trials were presented in randomized order). We asked participants to imagine that these partners were acquaintances, each a different one (they were not specifically asked to have real-world acquaintances in mind). We used this data to infer, for each participant, this participant's prior belief about a partner's WTR. We averaged these priors to generate a prior for the ideal observer (see SOM for details)[5].

[5] We use this prior for the analyses of both Studies 1 and 2 reported in the main text. In our pre-registration for Study 1, we pre-registered a different prior, based on existing empirical data about how people *do* play as dictators in the WTT. Although a model using this prior had a relatively good fit to the

Finally, the generative model used by the ideal observer features a parameter $\sigma_\phi$, quantifying the amount of noise that goes into people's welfare-tradeoff decisions (see above). We set this parameter's value by inferring the median value of $\sigma_\phi$ in an existing sample of participants playing the WTT as dictators (see SOM).

We compared human predictions to the predictions made by the ideal observer in two studies which used the prediction task described above. The ideal observer model, as well as the design of studies 1 and 2, were pre-registered[6], and the studies were approved by the Institutional Review Board at [Redacted for blind review]. The data, the R code for the computational model, data analyses, and figures are available at https://osf.io/3syce/?view_only=d5f639b7bda84962a72b83baa484b797.

## Study 1

### Participants

We recruited 100 US residents (40 female, mean age: 34.11) from Amazon MechanicalTurk. Following our pre-registration, we excluded 37 participants who failed an attention check, yielding a final sample of 63 (26 female, mean age: 34.86). We chose this sample size because it is very large, given the large number of trials per participant and the within-subjects nature of our main tests.[7]

---

data (we report this analysis in the SOM), we find that the model fit is improved by using a prior that is directly inferred from participants' prior judgments *about others*. We note that although we tested two different priors, our analysis is more conservative than the commonly used modeling practice of directly fitting parameters to the data with techniques like maximum likelihood estimation (see Lewandowsky & Farrell, 2010), which select the best-fitting values for a parameter among a wide array of initial candidates.

[6] https://osf.io/y8hks/?view_only=9948bf341a3d4c5d8df4aaf0d9baabd4. We address small deviations from the pre-registration in the SOM.

[7] For examples of recent studies testing Bayesian models with sample sizes in a similar range, see (Lopez-Brau et al., 2022; Marchant et al., 2023; Gong et al., 2023).

## Results

### *Do human predictions match ideal observer predictions?*

Yes. The item-level correlation between the average human prediction for a given trial and the model prediction for that trial was r(48)= .978, p <.001. Figure 2 shows that both human and model predictions are regulated by the same factors: Partners for whom the ideal observer inferred a high WTR elicit more optimistic predictions, and prediction trials with a high cost of giving elicit less optimistic predictions.

Figure 2 also reveals slight differences between model and human predictions. Human predictions are less likely to take extreme values (near 0 and 100). This might reflect the fact that human probability judgments are biased toward moderate values (see Zhu et al., 2020), or regression to the mean caused by occasional random responding. Human predictions also show a slightly larger discontinuity (compared to model predictions) between 'selfish' and 'generous' partners (at the left and right of the dashed line, respectively).

Human predictions were also correlated with model predictions when analyzed at the individual level: The median correlation between an individual's predictions and the model predictions (across trials) was r(48) = .86; inter-quartile range = .82 to .91; see Figure 3 (left). We report individual-level analyses in more detail in the SOM.

In the remainder of this section we analyze the data with linear mixed models. We z-scored the predictor and outcome variables we used in our linear mixed models, so that b coefficients can be interpreted as effect sizes.

### *Can human predictions be explained as the result of simple heuristics rather than inferences?*

Maybe participants did not engage in social valuation inferences, but made predictions by simply registering whether their partner chose to Give or Take, and/or by making less optimistic predictions in prediction trials when $\pi_{\text{partner}}$ (the opportunity cost

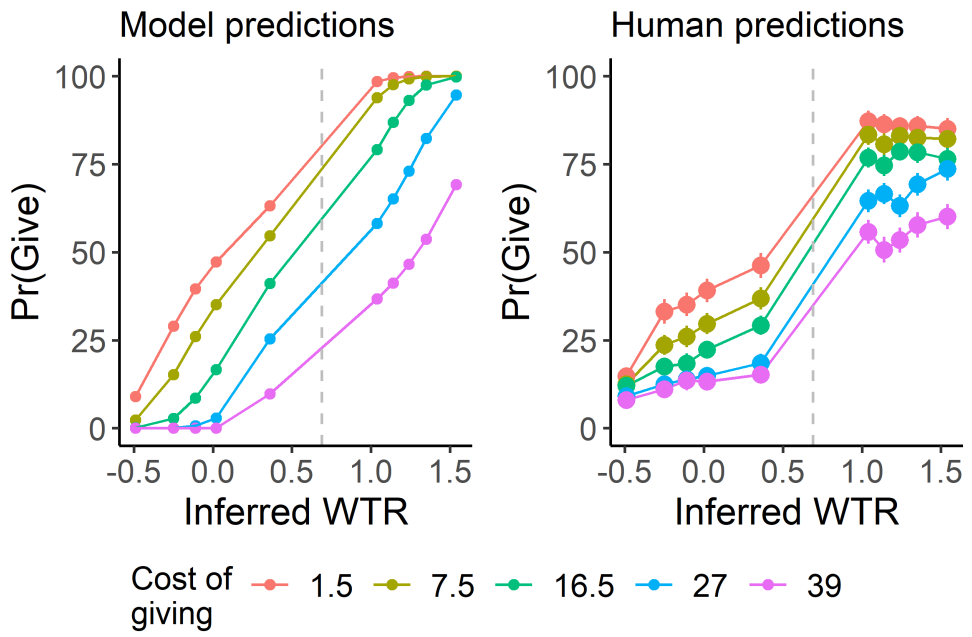


**Figure 2**

*Predictions made by the ideal observer* ***(left)*** *and average predictions made by human participants* ***(right)****, in Study 1. Each dot represents one prediction trial. In both panels, the x-axis represents the WTR that the ideal observer inferred that the partner had toward the participant. "Cost of giving" refers to* $\pi_{partner}$*: the potential payoff (in USD) for the dictator in that trial (while* $\pi_{participant}$ *was always $30 in each prediction trial). Error bars represent standard errors of the mean. Within each panel, selfish partners are left of the dashed line and generous partners are to its right.*

of giving) was large.

In order to rule out that possibility, we computed the association between the WTR that the ideal observer inferred the partner to have toward the participant, and participants' predictions for that partner. We did so while statistically controlling for a dummy variable indicating whether a partner was "selfish" (always took money for itself) or "generous" (always allocated the money to the participant). Henceforth, we refer to this dummy variable as "material payoffs", because each of the 5 selfish partners made decisions with the same aggregate material consequences, in terms of benefits gained and opportunity

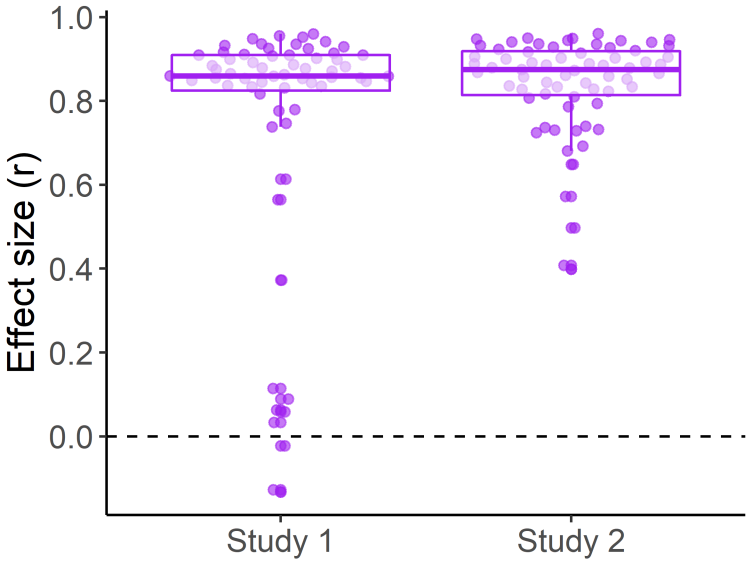


**Figure 3**

***Individual-level model fits, Study 1 (online sample) and 2 (undergraduate sample).*** *Each point corresponds to the Pearson's correlation coefficient (r) between model predictions and the predictions of one participant. Points are jittered along the x-axis for readability.*

costs inflicted (and similarly for the 5 generous partners). If participants did not make WTR inferences, but simply kept a tally of costs and benefits, they would make the same predictions for all 5 generous partners, and the same predictions for all 5 selfish partners.

Controlling for material payoffs, the WTR inferred by the ideal observer is positively associated with human predictions, $b = .33$, $p < .001$ (linear mixed model with random slopes and random intercepts, material payoffs and inferred WTR as fixed effects, and participant as a random effect). Therefore, simple heuristics based on material payoffs are insufficient to explain the data—participants appear to have been making inferences about social valuation.

***Does the WTR inferred by the ideal observer predict anger and gratitude?***

Yes for Anger, no for Gratitude. For the emotion analyses, we regress participants' emotion ratings about a partner on the WTR that the ideal observer inferred that partner to have.[8] The WTR inferred by the ideal observer was a negative predictor of Anger, b = -.53, $p < .001$, and a positive predictor of Gratitude, b = .82, $p < .001$ (linear mixed models with inferred WTR as fixed effect, random slopes and random intercepts, and participant as a random effect).

Controlling for material payoffs, ideal-observer-inferred WTR remained a significant predictor of Anger, b = -.44, $p < .001$, but it was no longer a significant predictor of Gratitude, b = .05, p = .27. In other words, participants' anger discriminated even among selfish partners: They were angrier toward those selfish partners who elicited lower WTR inferences in the ideal observer, even though each of the selfish partners inflicted the same overall opportunity cost on the participant ($35), and gained the same overall benefit ($60) by doing so (see figure 4). By contrast, variation in gratitude ratings was driven entirely by whether the partner had allocated money to themselves or to the participant.

Note that in both study 1 and study 2, Gratitude ratings for generous partners were near ceiling, with more than 50% of ratings being on the maximum point on the scale (7 on a 1-7 likert scale); this may have limited our ability to detect any effect of inferred WTR on Gratitude. Future studies could address this limitation in our design, for example by having participants play with partners who deliver smaller benefits.

In addition to our experimental evidence for a link between WTR inference and anger, we also find tentative correlational evidence. Participants who made pessimistic predictions about their partner (suggesting that they inferred a low WTR) tended to

[8] This test measures the effect of an experimental manipulation (the manipulation of the decisions made by different partners). One could alternatively use individualized estimates of a participant's inferences, but this would leave open the possibility that a third, unobserved variable is independently causing their WTR inferences and their anger.

report higher anger towards that partner. Even for participants observing the same partner, those reporting higher anger toward that partner subsequently were less likely to predict that the partner will allocate the money to the participant in prediction trials. This effect is slightly weaker, however, when controlling for a potential confound. We report these analyses in detail, for both Study 1 and 2, in the Supplementary Information.

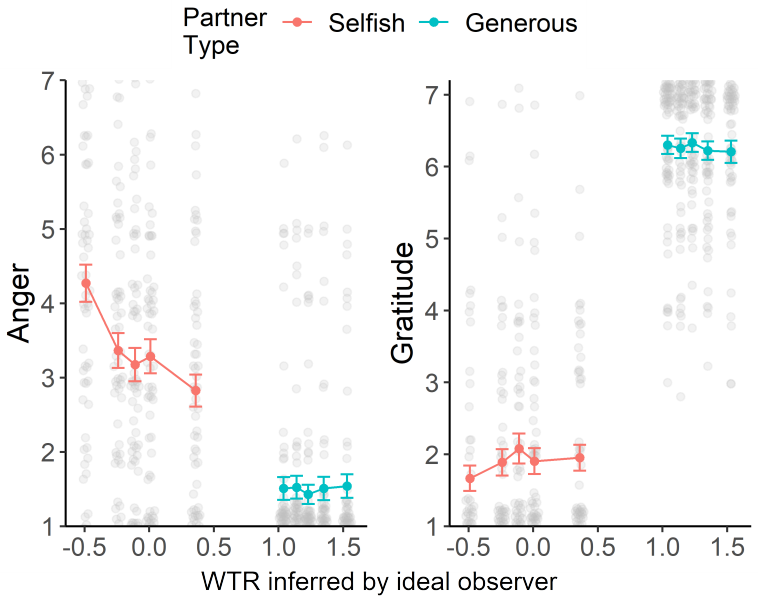


**Figure 4**

*Participants' mean anger and gratitude toward each partner as a function of the WTR inferred by the ideal observer for that partner, in Study 1. Grey dots represent individual ratings, and are jittered along the y-axis for visibility. Error bars represent standard errors of the mean.*

## Discussion

The predictions made by participants tracked the predictions made by the ideal observer model. This suggests that participants inferred the welfare-tradeoff ratio of their partners, and did so in an approximately rational way. Participants also expressed more anger at partners for whom the ideal observer inferred a low welfare-tradeoff ratio. In Study 2, we attempt to replicate these findings in a different sample. We also attempt to

directly measure participants' priors.

## Study 2

Study 2 was a replication of Study 1, with an additional phase at the beginning of the study. In this preliminary phase, participants were asked to make predictions about dictators for whom they had no information about past behavior in the Welfare Trade-off Task. We used their predictions in this phase to estimate the prior beliefs that participants have about the distribution of WTRs among their acquaintances. We then used these estimates to determine the prior of the ideal observer (see methods above). Study 2 also used an undergraduate sample instead of an online sample, because we thought an undergraduate sample would yield more precise individual-level data, given the more controlled environment of the laboratory.

We recruited 100 participants (72 female, 1 other, mean age: 18.8) from the undergraduate psychology subject pool at a university in California. Participants completed the study on a desktop computer while seated in a semi-private cubicle. One participant failed to complete the study because of computer error. Following our pre-registration, we excluded from analysis 32 participants who failed either a probability comprehension check (4 participants) and/or an attention check (29 participants), yielding a final sample of 67 participants (48 female, 1 other, mean age: 18.8).

### Results

#### *Baseline prediction task*

Figure 5 displays the predictions made by participants in the preliminary prediction task, where they made predictions about partners for whom they had observed no previous decisions. Participants predicted a lower probability that the partner would give as the opportunity cost of giving ($\pi_{\text{partner}}$) increased, b = -0.60, $p < .001$. At the individual level, this relationship was significant for 59 out of 67 participants.

We found the values of $\mu$ and $\sigma$ for the ideal observer that would result in the closest fit to participants' predictions in this task (we obtain $\mu = .55$, $\sigma = 1.01$; see Methods and Supplementary Information). The predictions of the ideal observer with this parameterization are shown in orange in Figure 5. They were highly correlated with mean participant predictions, $r(18) = .994$, $p < .001$. We use this same parameterization of the ideal observer to model our main task in both Study 1 and 2.[9]

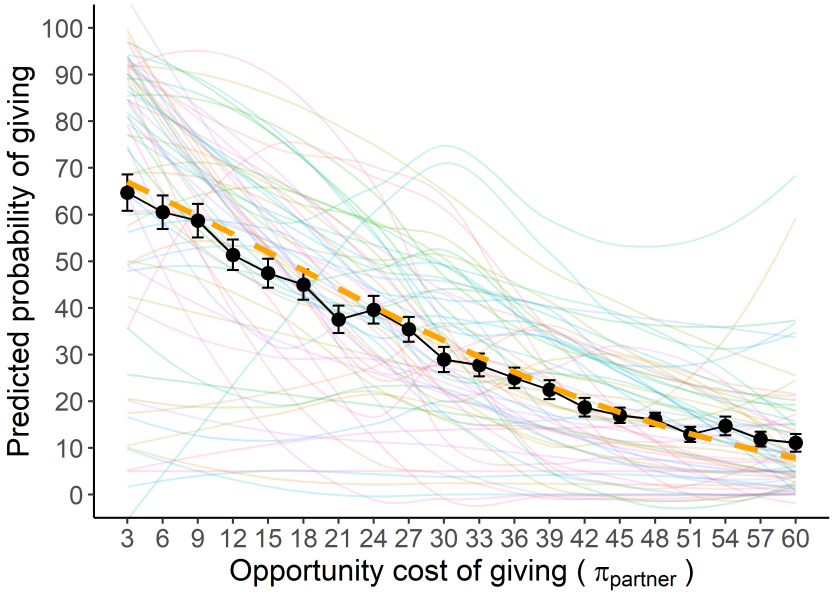


**Figure 5**

*Average human predictions (black), and best-fitting ideal observer predictions (dashed orange) in the baseline prediction task, Study 2. In each trial, $\pi_{participant}$ was* \$30*. Error bars represent standard errors of the mean. Faint colored lines represent loess regression fits for each individual participant.*

We now turn to results in the main task, where participants made judgments about partners for whom they had observed two previous decisions.

[9] The good fit of the model in the baseline prediction task might to some extent be attributed to over-fitting; the true test of the model lies in its ability to predict human judgments in the main task, without having to re-fit its parameters.

***Do human predictions match ideal observer predictions?***

Yes. The item-level correlation between the average human prediction for a given trial and the ideal observer's prediction for that trial was very large, r(48)= .988, p <.001; see Figure 6. Human predictions also correlated highly with the model predictions in individual-level analysis (see Figure 3, and the SOM).

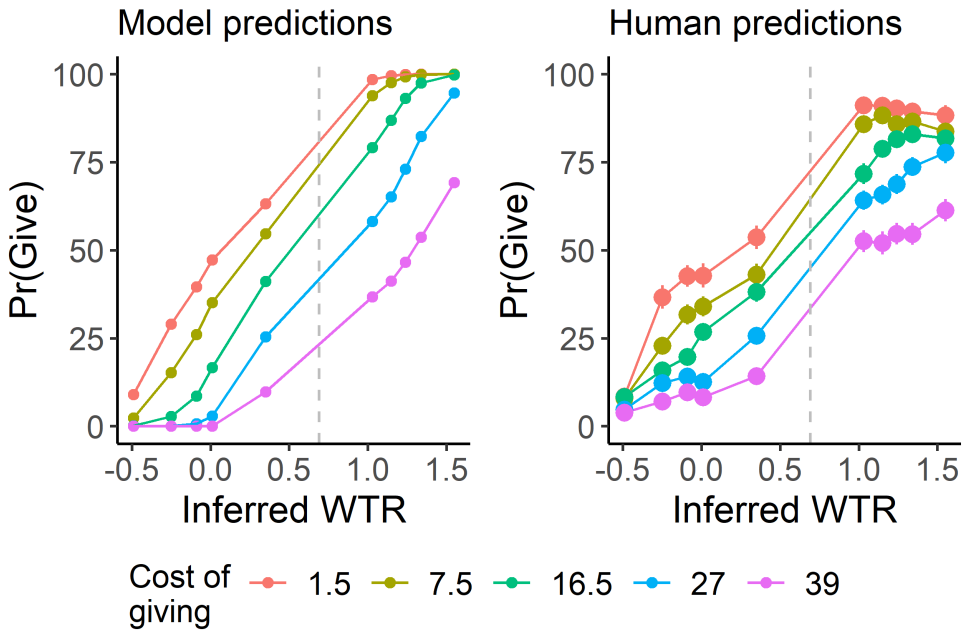


**Figure 6**

*Predictions made by the ideal observer* ***(left)*** *and average predictions made by human participants in Study 2* ***(right)****. Each dot represents one prediction trial. In both panels, the x-axis represents the WTR that the ideal observer inferred the partner to have toward the participant. "Cost of giving" refers to* $\pi_{partner}$*: the potential payoff (in USD) for the dictator in that trial (*$\pi_{participant}$ *was $30 in each prediction trial). Error bars represent standard errors of the mean. Within each panel, selfish partners are on the left of the dashed line and generous partners are on the right.*

***Can this result be explained by simple heuristics?***

No. Controlling for material payoffs, the WTR inferred by the ideal observer was positively associated with human predictions, b = .52, p < .001; (linear mixed model with random slopes and random intercepts, material payoffs and inferred WTR as fixed effects, and participant as a random effect).

***Does the WTR inferred by the ideal observer predict anger and gratitude?***

Yes. Figure 7 displays Anger and Gratitude ratings of participants for each partner, as a function of the WTR inferred by the ideal observer for that partner. Inferred WTR was a negative predictor of Anger, b = -.65, p < .001, and a positive predictor of Gratitude, b = .86, p <.001 (linear mixed models with random slopes and random intercepts, inferred WTR as fixed effect, and participant as a random effect).

When controlling for material payoffs, inferred WTR remained a significant predictor of Anger (b = -.73, p < .001) and Gratitude (b = .10, p = .02). Gratitude ratings were driven primarily by material payoffs, however.

## General discussion

An emerging body of research suggests that humans represent the weight that someone else assigns to their welfare, and that these representations play a key role in social cognition. Here we report new evidence for this hypothesis, by studying how people predict the behavior of others in a welfare-tradeoff task. Participants made predictions that closely tracked those of a Bayesian ideal observer that (i) represents the weight that the decision-maker attaches to the participant, and (ii) updates this weight rationally, based on just two decisions (only one of which was potentially informative). Participants' predictions could not be explained by simple heuristics, such as extrapolating from the amount of money their partner gave or failed to give them in the past.

For each person they had to evaluate, participants could observe only two of their

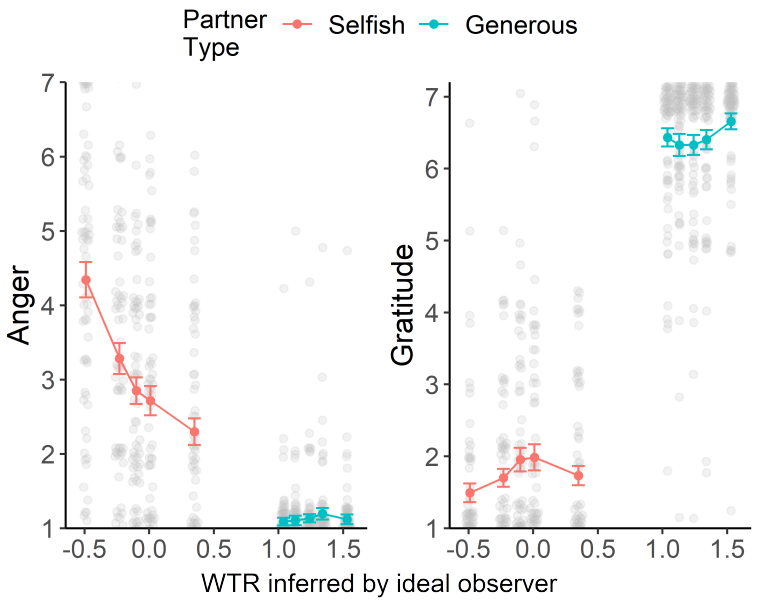


**Figure 7**

*Participants' mean anger and gratitude ratings for each partner in Study 2, as a function of the WTR inferred by the ideal observer for that partner. Grey dots represent individual ratings, and are jittered along the y-axis for visibility. Error bars represent standard errors of the mean.*

decisions. Often these decisions did not contain enough information to allow for straightforward predictions about how the person would behave in other contexts. Therefore, participants had to solve a difficult problem of statistical inference under uncertainty. The close fit between their predictions and those made by the ideal observer is surprising given the large body of work documenting that humans systematically deviate from normative statistical reasoning in many contexts (Kahneman et al., 1982; Marcus, 2009). On the other hand, our results provide additional evidence that, in ecologically valid contexts, human statistical inference can approximate Bayesian standards (Knill & Richards, 1996; Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Griffiths & Tenenbaum, 2006; Weiss et al., 2002).

Note that a domain-general ability to draw sound statistical inferences would not be enough, on its own, to generate the judgments that participants made. The ideal observer

model relies on a 'generative model': a set of domain-specific assumptions about the way people typically make welfare trade-offs. The generative model used by the ideal observer assumes that agents (noisily) maximize a utility function; this assumption is consistent with existing theories of how people reason about the preferences of others (Lucas et al., 2014; Jara-Ettinger et al., 2016; Jern et al., 2017; Baker et al., 2017). Hence our findings provide new empirical support for these theories.

In addition, the generative model assumes that the minds of other agents contain parameters that regulate the relative weight that the agent assigns to the welfare of the participant relative to its own. The tight fit between model and human behavior suggests that participants have inference systems that access a similar set of domain-specific assumptions. Thus, our findings support the proposal that humans represent the minds of other agents as containing welfare trade-off parameters (Tooby et al., 2008; Sell et al., 2017; Qi & Vul, 2022; Quillien, 2023).

**Implications for the recalibrational theory of anger**

Our findings also provide new empirical support for the recalibrational theory of anger (Sell et al., 2017; Sell et al., 2009; Sell, 2005). According the recalibrational theory, anger is an emotion designed to bargain with others for better treatment. It is activated by cues that the weight a person assigns to your welfare is too low. This inference motivates you to 'recalibrate' that weight by (for example) threatening to withdraw cooperation or inflict costs through aggression. Consistent with the theory, people with higher ability to confer costs and benefits (as indexed by physical formidability or attractiveness) get angry more easily (Sell et al., 2009), and the signature facial characteristics of anger enhance cues of strength (Sell et al., 2014). Apologies are an effective way to assuage anger when the apology signals that the target of anger did not intend to harm the offended party, or has resolved to put more weight on their welfare in future decisions (Sell et al., 2017).

A key prediction of the recalibrational theory is that we should get angry when we

infer that someone's valuation of our welfare is too low. As evidence for this prediction, prior research has found that inflicting a cost on a target elicits more anger when the offender did it to gain a trivial benefit rather than a large one (Sell et al., 2017; Sell, 2005). People are also angrier when they are specifically targeted by the offending action (Sell et al., 2017; Molho et al., 2017). Here we find more direct evidence that anger depends on an inference about social valuation: Partners who made decisions that imply a lower valuation of the participant (as assessed by the ideal observer) elicited more anger. Importantly, this was true even when holding constant the total opportunity costs that each partner inflicted on the participants, and the total gains that each partner obtained at the expense of the participants.

We found only weak evidence for an association between social valuation inference and gratitude, although this might be due to a ceiling effect for gratitude ratings. The extent to which gratitude depends on valuation inferences remains an important area for future research (see Forster et al., 2022).

## Limitations and directions for future research

### *Extensions of the experimental paradigm*

Many potential variations on our experimental design are possible. Letting participants observe more decisions per partner, for example, would allow us to understand the long-term dynamics of learning.[10] We could also have participants interact with partners who make *inconsistent* decisions. If your partner sometimes acts selfishly and sometimes generously, but in a way that is only weakly correlated with the cost of giving, it is difficult to make inferences about her welfare-tradeoff ratio. On the other hand, the

[10] As noted in the introduction, we focused on participants' ability to extrapolate from a few decisions because this is a stringent test of their ability to reason under uncertainty. As participants gather more evidence, we expect them to make predictions that are closer to 0 or 100% because they have less uncertainty about their partner's WTR.

level of consistency in your partner's decisions is something that you could, in principle, learn, given enough data. The ideal observer we used here does not attempt to learn the level of noise in an agent's decision-making, but this method could be extended to model such learning (by making joint inferences about an agent's WTR and consistency). Experimental tests could then probe to what extent participants' judgments about their partner's consistency match those of the ideal observer. We also conjecture that making welfare-tradeoffs in an incoherent manner might itself be a cue of poor valuation; it suggests that you are investing no cognitive resources in evaluating the consequences your decisions have on others (see Sarin & Cushman, 2022).

We used a simple experimental manipulation, only varying the payoff sets in the decisions that participants observed. This restricted parameter space made modeling more tractable, but it also over-simplifies the inference tasks that people are facing in the real world. For example, in many cases it is not clear to what extent a person's decision was caused by the fact that they value the decision's outcome — in folk-psychological language, whether the person did what they did 'intentionally' (Quillien & German, 2021). Uncertainty about intentionality adds a layer of complexity to the problem of inferring the weight someone assigns to your welfare: If your host serves a dish that contains an ingredient you are allergic to, is it because they do not care about you, because they forgot about your allergy, or because they do not know you are allergic? People are less angry at unintentional compared to intentional harm infliction (Sell et al., 2017), suggesting that they appropriately adjust their social valuation inferences to account for intentionality. Future research could investigate whether people appropriately factor *uncertainty* about intentionality when they make valuation inferences (see also Davis et al., 2021).

### *Third-party relationships*

Our experiment focused on the simple dyadic case, where the reasoner infers how much someone else values them. People can also make inferences about third-party

relationships, inferring for instance the relationship between A and B by seeing how A treats B. They also expect that the way that A treats B contains information about how A might treat them (Krasnow et al., 2016; Delton & Krasnow, 2017). These sophisticated inferences are important inputs to punishment and partner choice (e.g. Baumard et al., 2013; Krasnow et al., 2016), and the computational framework developed here is a potentially fruitful way to study them. Registering the patterns of social valuation among people in your social environment is also important to track group dynamics; future research should map whether and how welfare-tradeoff inferences can provide a foundation for coalitional cognition (Delton & Krasnow, 2017; Qi et al., 2022; Pietraszewski, 2022; Lindner & Krasnow, 2022).

### *Reasoning about the emotions of others*

Mapping the computational architecture of social emotions is key to understanding how people feel, but also how they reason about the emotions of *other people*. A core prediction of the recalibrational theory is that people will implicitly interpret anger as expressing a demand for better treatment, because verbal and nonverbal expressions of anger function as a signal (Sell et al., 2017; Sell et al., 2014). This perspective is convergent with an emerging body of work charting people's intuitive theory of emotions at a computational level (Ong et al., 2019; Houlihan et al., 2023). An intuitive theory of *social* emotions, in particular, allows one to jointly reason about how people will feel in a given situation, and the nature of the relevant social relationship (Smith-Flores & Powell, 2023). For example, people might expect Alice to be upset when Bob forgets about her birthday, but only if they know that Bob and Alice are close. Conversely, observing that Alice is upset at Bob's obliviousness might lead people to infer that Alice and Bob have a close relationship. Evidence suggests that these sorts of inferences appear early in development (Smith-Flores & Powell, 2023); future research can leverage computational theories to derive new fine-grained predictions in this domain.

### *Extensions of the model*

Our work involves important idealizations at the theoretical level. Our model effectively assumes that people track the value of a scalar variable in the mind of another person (a welfare-tradeoff ratio; for computational models making similar assumptions, see for example Qi & Vul, 2022; Jern & Kemp, 2014; Ullman et al., 2009; Davis et al., 2021). This is a valid approximation in contexts where an agent's utility function remains fixed across possible decisions. But evidence suggests that across different contexts, decision-makers construct utility functions on the spot, in a flexible and adaptive manner (Dana et al., 2007; Kleiman-Weiner, Shaw, et al., 2017; Bardsley, 2008; Guzmán et al., 2022). We expect that people adjust their inferences accordingly: for instance if someone was generous when their behavior was observed, people may not assume that this person will be equally generous when unobserved. Future work could profitably develop models of how people make these inferences, in a hierarchical Bayesian framework, for example (Kemp et al., 2007; Kleiman-Weiner, Saxe, et al., 2017).

### *Cross-cultural validity*

Finally, our work was conducted with participants from the United States. Previous research has found high cross-cultural convergence in the way that people make welfare-tradeoffs (Delton et al., 2023) and the way they respond to cues of social valuation (Sell et al., 2017). This prior research leads us to expect that the current results will generalize to participants in different cultures, including small-scale (non-WEIRD) societies. This prediction remains to be tested.

## Conclusion

The current work provides evidence that people infer how much someone else values their welfare, and that these inferences exhibit the fingerprints of good functional design. The results bolster the idea that representations of social valuation play a key role in

human social life.

## References

Algoe, S. B. (2012). Find, remind, and bind: The functions of gratitude in everyday relationships. *Social and personality psychology compass*, *6*(6), 455–469.

Andreoni, J., & Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396.

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.

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior*, *34*(3), 164–175.

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental economics*, *11*(2), 122–133.

Barnby, J. M., Bell, V., Mehta, M. A., & Moutoussis, M. (2020). Reduction in social learning and increased policy uncertainty about harmful intent is associated with pre-existing paranoid beliefs: Evidence from modelling a modified serial dictator game. *PLoS computational biology*, *16*(10), e1008372.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59–78.

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. (2022). How inferred motives shape moral judgements. *Nature Reviews Psychology*, *1*(8), 468–478.

Cosmides, L., Barrett, H. C., & Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences*, *107*(supplement_2), 9007–9014.

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

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1–73.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80.

Davis, I., Carlson, R. W., Dunham, Y., & Jara-Ettinger, J. (2021). Reasoning about social preferences with uncertain beliefs. *PsyArXiv*.

Debove, S., André, J.-B., & Baumard, N. (2015). Partner choice creates fairness in humans. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1808), 20150392.

Delton, A. W., Cosmides, L., Guemo, M., Robertson, T. E., & Tooby, J. (2012). The psychosemantics of free riding: Dissecting the architecture of a moral concept. *Journal of personality and social psychology*, *102*(6), 1252.

Delton, A. W., Jaeggi, A. V., Lim, J., Sznycer, D., Gurven, M., Robertson, T. E., Sugiyama, L. S., Cosmides, L., & Tooby, J. (2023). Cognitive foundations for helping and harming others: Making welfare tradeoffs in industrialized and small-scale societies. *Evolution and Human Behavior*.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, *38*(6), 734–743.

Delton, A. W., & Robertson, T. E. (2012). The social cognition of social foraging: Partner selection by underlying valuation. *Evolution and human behavior*, *33*(6), 715–725.

Delton, A. W., & Robertson, T. E. (2016). How the mind makes welfare tradeoffs: Evolution, computation, and emotion. *Current Opinion in Psychology*, *7*, 12–16.

Delton, A. W. (2010). *A psychological calculus for welfare tradeoffs*. University of California, Santa Barbara.

Dhaliwal, N. A., Martin, J. W., Barclay, P., & Young, L. L. (2022). Signaling benefits of partner choice decisions. *Journal of Experimental Psychology: General*, *151*(6), 1446.

Eisenbruch, A. B., & Krasnow, M. M. (2022). Why warmth matters more than competence: A new evolutionary approach. *Perspectives on Psychological Science*, 17456916211071087.

Eisenbruch, A. B., & Roney, J. R. (2017). The skillful and the stingy: Partner choice decisions and fairness intuitions suggest human adaptation for a biological market of cooperators. *Evolutionary Psychological Science*, *3*(4), 364–378.

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.

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, *97*(5), 1858–1876.

Forster, D. E., Pedersen, E. J., McCullough, M. E., & Lieberman, D. (2022). Evaluating benefits, costs, and social value as predictors of gratitude. *Psychological Science*, *33*(4), 538–549.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision research*, *51*(7), 771–781.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological review*, *102*(4), 684.

Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (2023). Active causal structure learning in continuous time. *Cognitive Psychology*, *140*, 101542.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767–773.

Guzmán, R. A., Barbato, M. T., Sznycer, D., & Cosmides, L. (2022). A moral trade-off system produces intuitive judgments that are rational and coherent and strike a balance between conflicting moral values. *Proceedings of the National Academy of Sciences*, *119*(42), e2214005119.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235.

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, *88*, 103948.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of theoretical biology*, *7*(1), 17–52.

Hammerstein, P., & Parker, G. A. (1982). The asymmetric war of attrition. *Journal of Theoretical Biology*, *96*(4), 647–682.

Helmholtz, H. (1856). *Treatise on physiological optics* (Vol. 3). Courier Corporation.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences*, *112*(6), 1727–1732.

Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B., & Saxe, R. (2023). Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220047.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naıve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

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

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

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, *10*(3), 307–321.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? *Proceedings of the Cognitive Science Society*.

Knill, D. C., & Richards, W. (1996). *Perception as bayesian inference*. Cambridge University Press.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological science*, *27*(3), 405–418.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE publications.

Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature*, *445*(7129), 727–731.

Lim, J. (2012). *Welfare tradeoff ratios and emotions: Psychological foundations of human reciprocity*. University of California, Santa Barbara.

Lindner, M., & Krasnow, M. (2022). Indirect intergroup bargaining: An evolutionary psychological theory of microaggression. *Evolutionary Psychological Science*, *8*(4), 478–492.

Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2022). Social inferences from physical evidence via bayesian event reconstruction. *Journal of Experimental Psychology: General*.

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

Marchant, N., Quillien, T., & Chaigneau, S. E. (2023). A context-dependent bayesian account for causal-based categorization. *Cognitive Science*, *47*(1), e13240.

Marcus, G. (2009). *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt.

Maynard-Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*(5427), 15–18.

Molho, C., Tybur, J. M., Güler, E., Balliet, D., & Hofmann, W. (2017). Disgust and anger relate to different aggressive responses to moral violations. *Psychological science*, *28*(5), 609–619.

Monroe, A. (2020). Moral elevation: Indications of functional integration with welfare trade-off calibration and estimation mechanisms. *Evolution and Human Behavior*, *41*(4), 293–302.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *science*, *314*(5805), 1560–1563.

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, *11*(2), 338–357.

Pietraszewski, D. (2022). Toward a computational theory of social groups: A finite set of cognitive primitives for representing any and all social groups in the context of conflict. *Behavioral and brain sciences*, *45*, e97.

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

Qi, W., & Vul, E. (2022). The evolution of theory of mind on welfare tradeoff ratios. *Evolution and Human Behavior*.

Qi, W., Vul, E., Schachner, A., & Powell, L. J. (2022). Triadic conflict "primitives" can be reduced to welfare trade-off ratios. *Behavioral and Brain Sciences*.

Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of theoretical biology*, *492*, 110204.

Quillien, T. (2023). Rational information search in welfare-tradeoff cognition. *Cognition*.

Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition*, *214*, 104806.

Raihani, N. J., & Barclay, P. (2016). Exploring the trade-off between quality and fairness in human partner choice. *Royal Society open science*, *3*(11), 160510.

Sarin, A., & Cushman, F. (2022). One thought too few: Why we punish negligence.

Schmid, L., Chatterjee, K., Hilbe, C., & Nowak, M. A. (2021). A unified framework of direct and indirect reciprocity. *Nature Human Behaviour*, *5*(10), 1292–1302.

Sell, A. (2005). *Regulating welfare tradeoff ratios: Three tests of an evolutionary-computational model of human anger*. University of California, Santa Barbara.

Sell, A., Cosmides, L., & Tooby, J. (2014). The human anger face evolved to enhance cues of strength. *Evolution and Human Behavior*, *35*(5), 425–429.

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., Rascanu, R., Sugiyama, L., Cosmides, L., & Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, *168*, 110–128.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*(35), 15073–15078.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature human behaviour*, *2*(10), 750–756.

Smith, A., Pedersen, E. J., Forster, D. E., McCullough, M. E., & Lieberman, D. (2017). Cooperation: The roles of interpersonal value and gratitude. *Evolution and Human Behavior*, *38*(6), 695–703.

Smith-Flores, A. S., & Powell, L. J. (2023). Joint reasoning about social affiliation and emotion. *Nature Reviews Psychology*, 1–10.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sznycer, D., Delton, A. W., Robertson, T. E., Cosmides, L., & Tooby, J. (2019). The ecological rationality of helping others: Potential helpers integrate cues of recipients' need and willingness to sacrifice. *Evolution and Human Behavior*, *40*(1), 34–45.

Sznycer, D., Sell, A., & Lieberman, D. (2021). Forms and functions of the social emotions. *Current Directions in Psychological Science*, *30*(4), 292–299.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. *Handbook of approach and avoidance motivation*, *15*, 251.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, *46*(1), 35–57.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, *22*.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature neuroscience*, *5*(6), 598–604.

Woo, B. M., & Spelke, E. S. (2023). Toddlers' social evaluations of agents who act on false beliefs. *Developmental Science*, *26*(2), e13314.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*(3), 1099–1108.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological review*, *127*(5), 719.