When giving more makes you look worse: paradoxical inferences in a Bayesian model of social evaluation

Madeleine Horner[1*], Victor Btesh[2*], Adam Moore[1] & Tadeg Quillien[1]

[1] University of Edinburgh, Edinburgh, Scotland

[2] University College London, London, England

[*] Equal contribution: m.horner-1@ed.ac.uk;victor.btesh.19@ucl.ac.uk

Abstract

People readily infer how much another agent cares about their welfare, for example after observing this agent give them some of what they have. However, these inferences become more difficult when there is uncertainty over the resources someone has to share, a common real-world scenario. We develop a Bayesian computational model of how people infer the welfare trade-off ratio (WTR) of another agent under resource uncertainty. This model predicts that under uncertainty people should average over different possible hypotheses about their partner's resources. In a behavioral study (N = 129), we found that across donation amounts participants' WTR estimates closely tracked those made by our Bayesian model. Notably, participants exhibited the following paradoxical pattern predicted by our model: they sometimes saw agents as *less* generous when they gave them *more* money, in cases where a high donation revealed that the agent is rich and is giving a comparatively small portion of their resources. These findings demonstrate that people are able to make sophisticated inferences consistent with rational Bayesian reasoning, updating their beliefs about a partner's resources and adjusting their WTR estimates accordingly.

Keywords: Welfare Trade-off Ratio; Bayesian modeling; Inference; Social cognition

Introduction

A key challenge in social life is to evaluate other people according to important dimensions. It is for example important to assess whether a potential partner is competent, or whether they adhere to the rules of social exchange (Fiske, Cuddy, & Glick, 2007; Mercier, Morin, Mercier, & Quillien, 2026; Cosmides, Barrett, & Tooby, 2010). A particularly important dimension on which we evaluate others is their generosity, or more generally, their disposition to value our welfare when making decisions. An emerging body of work shows that people are adept at evaluating someone's generosity, making rational inferences from even thin slices of behavior (Quillien, Tooby, & Cosmides, 2023).

In this paper, we probe this ability further; we ask whether people can rationally infer someone's generosity even when contextual uncertainty complicates the inference problem. In particular, we consider a resource division setting where the amount of resources at the disposal of one's partner is unknown. In this context, if our partner gives us very little, it might be because they do not value us, or simply because they did not have much to share in the first place. Making an appropriate inference about how much our partner values us requires taking into account this uncertainty. Here we test if people are able to account for uncertainty in a way that accords with Bayesian principles of rational inference.

Our setting has an interesting property: in some situations, sharing *more* of one's resources might lead one to be seen as *less* generous, at least to the eyes of a rational Bayesian observer. This 'paradoxical' inference arises because the amount of money someone gives is also a cue to the resources they have at their disposal. In our experiment, a partner sometimes gives a moderate amount of what they have, enough to reveal to a Bayesian observer that they are wealthy, but not enough to appear generous. In this instance the Bayesian observer evaluates the partner as less generous than if they had given less and maintained uncertainty about their wealth. If our participants make rational inferences, they should make the same paradoxical inference that is made by a Bayesian observer.

The psychology of social valuation

An important aspect of social life is being able to infer whether other people are disposed to care for our welfare. People often make decisions that affect both the welfare of themselves and the welfare of others (Delton & Robertson, 2016) and successful evaluation of how much someone values us is crucial for predicting and explaining their future behavior (Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008; Eisenbruch & Krasnow, 2022; Powell, 2022). While these social evaluations feel effortless, the computations underlying them are complex and involve causal and statistical inference under uncertainty (Quillien et al., 2023).

An agent's concern for the welfare of another agent is often formalized as a *welfare trade-off ratio*, or WTR (Delton

et al., 2023; Qi, Vul, & Powell, 2025). The welfare trade-off ratio is the weight that someone places on the welfare of another compared to their own (Cosmides & Tooby, 2013). For example, if Alice has a high WTR for Bob, she will avoid harming him for trivial gains to herself, and will help him even at a cost to herself.

An emerging body of research has shown that someone's welfare trade-off ratio toward ourselves is a key target of social evaluation. In simple economic games, when a participant perceives that a partner has a high WTR for them, they will choose to continue interacting with that partner over more productive alternative partners (Lim, 2012; Hackel, Mende-Siedlecki, & Amodio, 2020; Raihani & Barclay, 2016). Likewise, social emotions like anger and gratitude are triggered by the inference that someone has an unexpectedly low or high WTR (Sell et al., 2017; Quillien et al., 2023).

Social inference and inverse planning

How can people infer the welfare trade-off ratio of others? *Inverse planning* provides a formal framework for addressing this question. Inverse planning consists in making inferences about the internal states of others by working backwards from their observable behavior, using Bayesian inference (Baker, Saxe, & Tenenbaum, 2009; Wu, Baker, Tenenbaum, & Schulz, 2018; Jern, Lucas, & Kemp, 2017; Houlihan, Ong, Cusimano, & Saxe, 2022). People appear to have a robust expectation for others to act rationally (Eisenhardt & Zbaracki, 1992), and use this assumption to infer the beliefs and desires of another agent, even from sparse observations of the agent's behavior (Baker et al., 2009; Ullman et al., 2009; Lucas et al., 2014; Ong, Zaki, & Goodman, 2018; Jara-Ettinger, Schulz, & Tenenbaum, 2020).

To illustrate, Alice's decision to help Bob contains information about her WTR toward Bob. If helping is very costly, we should infer that Alice's WTR is high, because she would not have paid the cost of helping if she did not value Bob's welfare highly (Quillien et al., 2023; Horner, Quillien, & Moore, 2025). An emerging body of work suggests that people are able to infer an agent's WTR in a way that is consistent with rational inverse planning models (Ullman et al., 2009; Jern & Kemp, 2014; Davis, Carlson, Dunham, & Jara-Ettinger, 2023; Quillien et al., 2023; Btesh, Lagnado, & Gerstenberg, 2025). For example, people reliably infer that an agent who hurts another in order to gain a trivial benefit has a low WTR toward the other agent (Quillien et al., 2023; Sell et al., 2017).

However, in previous work participants were given rich information about the relevant situation, for example the payoffs associated with different actions available to the agent. In real life, people often lack complete information about the situational context in which others act. For example, it is not always transparent whether someone acted intentionally, or how costly it would have been for them to help. We know little about how people make social inferences in these more ecologically common scenarios where contextual information is scarce.

When relevant information is unavailable, Bayesian inference requires *marginalizing* over the uncertainty (Jaynes, 2003). That is, a rational observer should consider what would follow under the possible scenarios consistent with their current evidence, weighting these scenarios by their probability. Here we ask whether people can engage in this kind of sophisticated probabilistic reasoning when making social inferences. As a case study, we consider a setting where there is uncertainty about the wealth of the agent under evaluation. Specifically, an agent has the opportunity to share a pot of money with the participant, and the amount of this pot of money (the agent's *endowment*) is hidden. Participants have to infer how much the agent values them, on the basis of the amount the agent decides to share. Crucially, performing this inference rationally requires accounting for one's uncertainty about the agent's endowment.

Methods

Ethics and Open Science

We preregistered this study's procedure, hypotheses, and analysis plan on the Open Science Framework (see `https://osf.io/ujqhs/overview?view_only=7f7d4b237fbf4dcf8808bcc13e0ae832`). The data and code for both the modeling and analysis can be found here: `https://github.com/Vbtesh/cogsci2026_inferring-welfare-tradeoff-ratio`. This study received ethical approval (Approval number: 34-2526/2 at Edinburgh University) and follows all British Psychological Society guidelines.

Participants

Participants (N = 140) were recruited through Prolific. Eligibility criteria required them to be over 18, residents of the UK or United States, and fluent in English. Participants were excluded from analysis for failing one or more comprehension questions.[1] After exclusion, data were analyzed for 129 participants (71 Female, mean age = 46). Participants were compensated £1.05. All participants signed consent forms prior to participation.

Overview of the Task

Participants, in the role of the receiver, played a modified version of a dictator game using hypothetical money; see

[1] Comprehension Questions: (1) Is the ball received by your acquaintance randomly drawn? (2) Which of the following were the colors of the balls shown?

Figure 1. On each trial, participants were paired with a different fictitious actor. The actor was given some amount of money (their 'endowment'), and decided how to split this endowment with the receiver. However, across trials participants could not always see the endowment of the actor. Participants were shown the raw monetary amount that the actor shared with them. Participants then indicated how much they thought the actor cared about them based on how much money the actor shared.

Materials and Design

Participants were shown a series of situations in which an actor shared a hypothetical amount of money with them. Participants were told that the amount of money at the actor's disposal (the *endowment*) was determined by a random draw from a raffle machine. This raffle machine contained ten colored balls of two different types; the actor received $100 when drawing a high-value ball and $20 when drawing a low-value ball; see Figure 1.

We used a 2 (Prior) $\times$ 2 (Endowment) $\times$ 2 (Knowledge) $\times$ 6 (Donation) mixed experimental design. Between-subjects, we manipulated participants' Prior belief about the probability of the actor receiving either $100 or $20. Specifically, we manipulated the relative proportion of low- to high-value balls in the raffle machine: either 80% low-value balls (Low Prior condition) or 80% high-value balls (High prior condition). The ball color associated with each monetary amount was randomized between participants.

Within subjects, we manipulated the actor's Endowment (High vs Low), the participant's Knowledge (whether the endowment was visible or not), and the donation amount (0%, 15%, 25%, 50%, 75%, or 100% of the endowment). Note that participants were shown the raw monetary amount of the donation, not the percentage. Each trial corresponded to one combination of Endowment, Knowledge, and Donation. Each participant saw all possible combinations, for a total of 24 trials per participant. Trials were presented in random order.

Procedure

Before starting the experiment, participants completed an example trial. Subsequently, participants completed a total of 24 trials. Participants were instructed to treat each situation as independent; to encourage this each trial had uniquely named actors.

In a given trial, participants were first shown the raffle machine from which the actor randomly received a ball. Second, in the Known endowment condition, participants were shown which ball the actor received and the corresponding endowment; in the Unknown endowment condition they were shown a gray ball and told they did not know the actor's endowment. Participants were then shown the raw monetary amount the actor decided to share with them. After receiving all these pieces of information, participants were asked to provide judgments about the welfare trade-off ratio of the actor ('How much do you think the actor cares about you?') and the empathy of the actor toward the participant ('How much empathy do you think the actor has for you?'). These judgments were collected using a 7-pt Likert Scale (1 = Not at all, 7 = Very Much). The order of these two questions was randomized between participants. The study was implemented using jsPsych (De Leeuw, 2015).

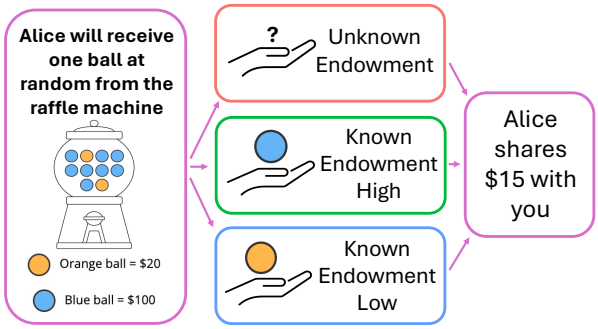

Figure 1: Structure of a trial in our experiment. The actor draws a ball from a raffle machine containing balls of different colors; the actor's endowment ($20 or $100) is determined by the color of the ball they draw. Participants were either given information about which ball the agent received or were given no information. Finally, participants were told how much money the agent shared with them.

Computational Model

In a task where an *actor* receives an endowment and can share some amount with a *recipient*, how can a Bayesian observer infer the welfare trade-off ratio of the actor? We assume that the actor makes decisions according to the following utility function:

$$U(a;\lambda,E) = f(E-a) + \lambda f(a) \quad (1)$$

Where E is the endowment of the actor, a is the donation amount, λ is the actor's welfare trade-off ratio toward the recipient, and $f(\cdot)$ is a function mapping money to agent-specific utility. In the pre-registration $f(\cdot)$ was simply the identity function $f(x) = x$. After data collection we also tested the function $f(x) = \log(x)$, which assumes diminishing marginal returns. This version of the model fit the data better, and is also psychologically more realistic (Kahneman & Tversky, 1979), so it is the one we report below. Technically we use $f(x) = \log(x + E \times \varepsilon)$ to avoid undefined outputs at $\log(0)$, with ε a free parameter.

The actor selects a donation amount a according to a softmax with inverse temperature τ^{-1} over the utility function:

$$p(a|\lambda, E) \propto \exp\{\tau^{-1} \cdot U(a;\lambda, E)\} \quad (2)$$

Given observed donation a, the observer can update their beliefs to obtain a joint distribution over λ and E using Bayes' rule:

$$p(\lambda, E|a) \propto p(a|\lambda, E)p(\lambda, E) \quad (3)$$

Finally the posterior $p(\lambda|a)$ can be derived by marginalizing over E:

$$p(\lambda|a) = \sum_{e \in E} p(\lambda, E = e|a) \quad (4)$$

We assume that λ and E are independent, such that $p(\lambda, E) = p(\lambda)p(E)$. The prior $p(\lambda)$ is a normal distribution with mean μ_λ and standard deviation σ_λ. The prior $p(E)$ depends on the experimental condition: in the Known Poor condition we have $p(E = 20) = 1$, in the Known Rich condition $p(E = 100) = 1$, and in the Unknown condition $p(E)$ assigns non-zero probability to both $E = 20$ and $E = 100$.

Auxiliary assumptions

Equations 1-4 define a rational Bayesian observer for our task. For comparison with the human data, we make two auxiliary assumptions. First, we allow for the possibility that participants do not perfectly encode the prior over the endowment. In the Unknown endowment condition, the prior $p(E)$ is in principle given by the relative proportion of high-value and low-value balls in the urn. To allow for the possibility that participants fail to perfectly encode these proportions, we assume that their prior $p(E)$ is a non-linear function of the objective probability $q(E)$:

$$p(E) = w\big(q(E); \alpha_{\text{prior}}, \beta_{\text{prior}}\big) \quad (5)$$

$$= \exp\{-\beta_{\text{prior}}\{-\log q(E)\}^{\alpha_{\text{prior}}}\} \quad (6)$$

with α_{prior} a free parameter controlling the degree to which the encoded probability deviates from the true proportion, and $\beta_{\text{prior}} = \log(2)^{1-\alpha_{\text{prior}}}$ is a deterministic function of α_{prior} that ensures the probability $1/2$ is mapped to itself.[2]

Second, to map the model's inference onto the 1-7 Likert scale used by participants, we use an ordered logit with 6 cutpoints. Altogether, the model has 5 free parameters and 6 bin cutpoints that we fit to the data (see Table 1). All aspects of the model, except the log utility function and the mapping between inferred λ and Likert scale ratings, were pre-registered.

[2]The α_{prior} parameter determines whether probabilities are underweighted below $1/2$ and over-weighted otherwise (when $\alpha_{\text{prior}} > 1$), or vice-versa ($\alpha_{\text{prior}} > 1$), see Prelec (1998).

Alternative models

As a baseline for comparison, we also fit two naive models (not pre-registered) which simply consider λ to be a monotone function of the donation amount. Under the first model, λ is an affine function of the donation amount; the second model works similarly but then passes the output of the affine function to a $\log(\cdot)$ function. The first model has 2 free parameters, a slope and intercept, while the second model has 3 free parameters, a slope and intercept as well as an ε parameter preventing the evaluations of $\log(0)$, analogous to the main model. The slope and intercept are both assumed to be Gaussian—therefore the models produce a Gaussian distribution over λ. To map λ to the 7-point Likert scale, we fitted the same ordered logit as the main model.

Results

We performed analyses with lme4 (Bates, Mächler, Bolker, & Walker, 2015) in R (version 4.3.2). All multilevel models are reported using standardized β coefficients. Degrees of freedom for coefficient tests have been obtained using Satterthwaite approximation in the lmerTest package (Kuznetsova, Brockhoff, & B.Christensen, 2017). The computational model was fit with the NumPyro probabilistic programming library in Python (Phan, Pradhan, & Jankowiak, 2019), using MCMC sampling.[3]

Modeling Results

Predictions of the Bayesian model were highly correlated with average human judgments across trials, $r(46) = .96$, $p < .001$. The Bayesian model also fit the data better than the alternative models, as assessed by their ELPD WAIC (Full model: -745.3 ($SE = 59.6$), affine model: -1726.3 ($SE = 198.0$), affine-log model: -1128.9 ($SE = 128.8$)). Posterior estimates for the full model are summarized in Table 1.[4][5]

This result suggests that participants' inferences approximated normative Bayesian updating. There was however one aspect in which participants deviated from the Bayesian ideal: they were not sensitive enough to the manipulation of the prior $p(E)$ over the endowment. Specifically, the best-fitting value of α_{prior} is very low, indicating that participants were under-sensitive to the difference in

[3]Convergence was assessed via the Gelman-Rubin statistic ($\hat{R}$) and effective sample size (ESS). All parameters achieved $\hat{R} = 1.0$ and sufficient ESS (bulk ESS > 1300, tail ESS > 800), indicating adequate convergence.

[4]The fitted bins corresponding to each points in the Likert scale are: $[-1, -0.21]$, $[-0.21, 0.23]$, $[0.23, 0.46]$, $[0.46, 0.71]$, $[0.71, 0.93]$, $[0.93, 1.21]$, $[1.21, 3.00]$.

[5]The best-fitting value of μ_λ is negative, in contrast to previous research (Quillien et al., 2023; Quillien, 2023). We suspect that this is a peculiarity of the optimization process and not a psychologically meaningful result. As a robustness check, we find that the model fits the data almost as well if we force μ_λ to be positive.

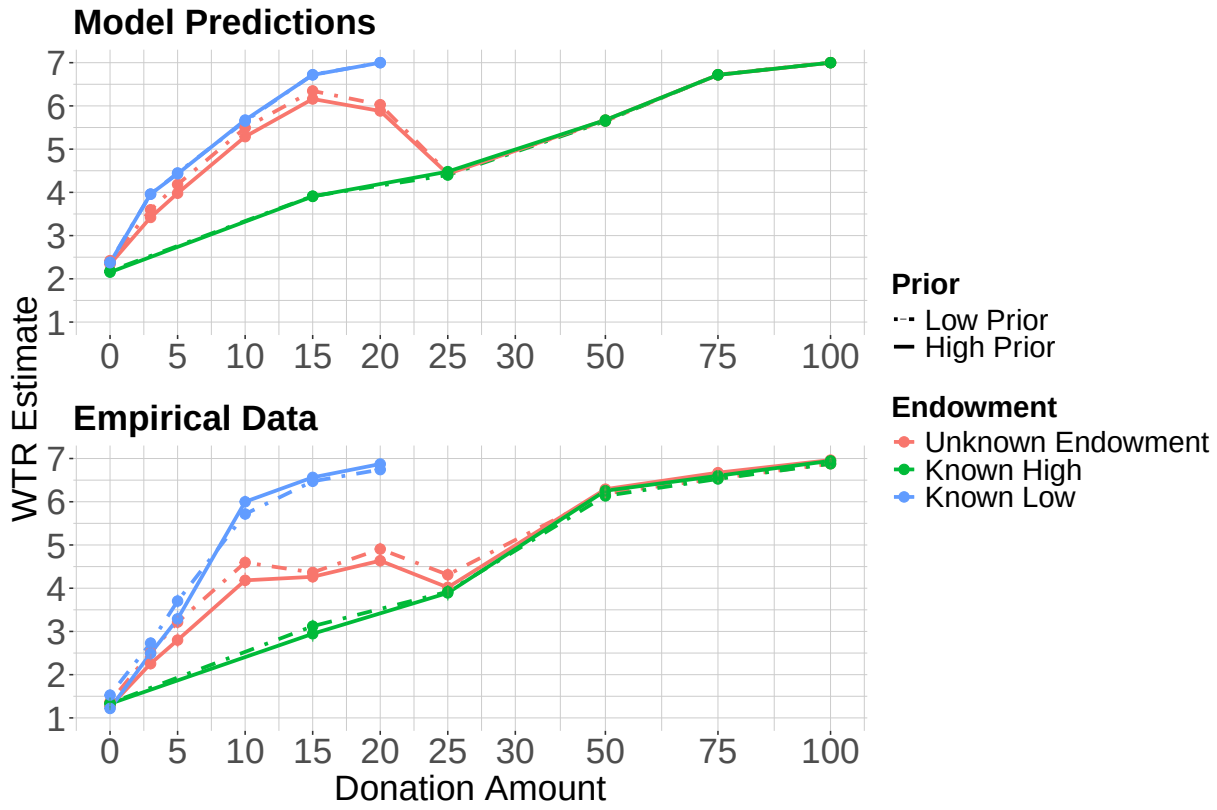

Figure 2: Model-predicted welfare trade-off ratio estimates and mean participant welfare trade-off ratio estimates as a function of donation amount and experimental condition. Line type (solid vs dotted) represents the prior over endowment. Line color represents what the participant knows about the actor's endowment. Note that for better legibility the spacing of the x-axis has been compressed for the donation amounts exceeding $30.

raffle machine composition between the Low and the High Prior condition. While the true prior probability of the actor having a high endowment was 0.2 in the Low Prior and 0.8 in the High Prior conditions, the best-fitting value of α_{prior} is consistent with participants assigning probabilities of 0.41 (95% HDI: [0.35, 47]) and 0.61 (95% HDI: [0.54, 0.67]) in the Low and High Prior conditions, respectively. In sum, participants exhibited a form of base-rate neglect, a widely documented bias in probabilistic reasoning (Kahneman & Tversky, 1973; Stengård, Juslin, Hahn, & Van den Berg, 2022).

Next we turn to a qualitative comparison between model predictions and participants' judgments.

Behavioral Results

Our computational model makes several key predictions about how people infer an actor's welfare trade-off ratio under uncertainty.[6]

Table 1: Posterior parameter estimates from MCMC sampling. Values represent posterior means with 95% highest density intervals (HDI).

Parameter	Mean (SD)	95% HDI
τ^{-1}	123.8 (20.57)	[87.6, 163.1]
μ_λ	-0.55 (0.07)	[-0.67, -0.42]
σ_λ	0.57 (0.02)	[0.53, 0.605]
ε	0.73 (0.05)	[0.64, 0.82]
α_{prior}	0.23 (1.04)	[0.10, 0.47]

Most critically, the model predicts a paradoxical pattern, whereby giving *more* money can sometimes make the actor appear *less* generous. Specifically, when the endowment is hidden, giving more than $20 reveals that the actor is rich (because an actor with the low endowment cannot give more than what they have). Giving $25, in particular, reveals both that the actor is rich and that they are sharing

[6]The statistical analyses reported in this section do not perfectly follow our pre-registration; this is because of space constraints and because the pre-registration omitted some key tests. Running all tests from the pre-registration leads to identical conclusions.

only one quarter of their endowment. In contrast, if the actor gives \$20, there is uncertainty about whether the actor is poor (making \$20 very generous) or rich (making \$20 relatively stingy). Therefore there is a counterintuitive *dip* in WTR estimates in the Unknown condition, from \$20 to \$25, that does not appear when the endowment of the actor is known (see Figure 2, upper panel).

We find robust evidence for this paradoxical dip in the human data (Figure 2, lower panel). In the unknown endowment condition (red line), WTR estimates were significantly higher for \$20 than for \$25 donations ($\beta = -0.121$, $SE = .02$, $t(128) = -5.57$, $p < .001$, mixed-effects regression with participant-level random effects). In contrast, when the endowment is known (blue and green lines), WTR estimates are a monotonically increasing function of donation amounts (both for model and participants).

The second key prediction of the model is that participants' judgments in the Unknown condition should be below their judgments in the known low condition but above their judgments in the known high condition (see upper panel in Figure 2, where the red line is below the blue line and above the green line). This prediction follows from the logic of *marginalization*: in the Unknown condition, a Bayesian observer must average between two possibilities: the actor might be Poor or Rich. For a given donation amount, a Poor actor should appear more generous than a Rich actor because they are sharing more of what they have. An actor whose endowment is Unknown should elicit an inference that falls in between these two cases. Importantly, this effect disappears for donation amounts above \$20, where a rational observer can infer that the actor is Rich even in the Unknown endowment condition.

To test this prediction, we ran two separate mixed-effects models comparing the Unknown endowment condition to each Known condition, restricting our analysis to donation amounts below \$25 (when the true endowment of the actor still remains ambiguous). As predicted, WTR estimates for the Known Poor endowment were significantly higher than in the Unknown endowment condition ($\beta = .293$, $SE = .093$, $t(2192) = 3.157$, $p < .001$) and estimates for the Known Rich endowment were significantly lower than in the Unknown endowment condition ($\beta = -1.05$, $SE = .108$, $t(1160) = -9.764$, $p < .001$).

The model also predicts higher WTR estimates when the actor is a priori more likely to be poor (i.e. when the proportion of low-value balls in the raffle machine is high). Furthermore, this effect is limited to the Unknown endowment condition, because when the endowment is visible prior information should no longer matter. Visually, the red dotted line is above the red solid line (upper panel on Figure 2). However, we did not find robust evidence for the predicted effect in the human data, although the effect is descriptively in the expected direction (see lower panel in Figure 2). In the Unknown condition, WTR estimates were not reliably lower in the Low compared to the High prior condition ($\beta = -0.09$, $SE = 0.10$, $t(127) = -0.93$, $p = .35$, mixed-effects model with participant-level random intercept).

Across all trials, judgments about the actor's empathy and welfare trade-off ratio estimates were highly correlated ($r = .97$, $p < .001$, 95% CI $[.97, .97]$). This finding is consistent with previous work highlighting the relationship between people's perceptions of the welfare trade-off ratio and the intuitive concept of empathy (Horner et al., 2025).

Discussion

People monitor whether other people care about their welfare, and can assess this even from sparse data (Sell et al., 2017; Davis et al., 2023; Quillien, 2023; Quillien et al., 2023). Here we probed this ability further, and explored whether people make rational inferences about an agent's welfare trade-off ratio (WTR) under contextual uncertainty. In everyday life, the resources of potential partners are not always known, which complicates the task of assessing what their actions reveal about their WTR. We find that participants make inferences that are consistent with Bayesian updating in this context, rationally accounting for their uncertainty about the amount of resources their partner has to share.

A Bayesian observer model makes a noteworthy prediction in our setting. In some situations, agents that share *more* money should be evaluated as *less* generous, because a large donation reveals the extent of one's wealth. Participants made the same paradoxical inference predicted by the model in these situations. Note that we are not arguing that this 'more is less' effect will always arise whenever observers have to jointly infer wealth and generosity. The 'more is less' effect arises in our task because the prior over the agent's endowment is a categorical distribution (the agent has either \$20 or \$100). We used the 'more is less' pattern as a test of whether people can make rational Bayesian inferences even when they are *prima facie* counter-intuitive.

Finally, the way people make social inferences has implications for how people make social decisions in the first place. Evolutionary considerations suggest, and empirical research has found, that human generosity is partly motivated by the goal of convincing others that we are good interaction partners (Roberts, 1998; Hardy & Van Vugt, 2006; Bardsley, 2008; Kleiman-Weiner, Shaw, & Tenenbaum, 2017; Btesh et al., 2025). From that perspective, our results show that decision-makers might sometimes be better off choosing the less generous of two possible allocations, even if they do not care about their own payoff and are only concerned with appearing generous. Future research could explore whether people sometimes act in this way.

Acknowledgments

We thank the reviewers for helpful feedback on a previous version of this manuscript. VB was supported by a Bogue Fellowship and an UCL Experimental Psychology Demonstratorship.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. doi: 10.1016/j.cognition.2009.07.005

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Btesh, V., Lagnado, D., & Gerstenberg, T. (2025). Taking others for granted: balancing personal and presentational goals in action selection. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*. Retrieved from https://escholarship.org/uc/item/23d145sd

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. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology*, *64*(1), 201–229. Retrieved from https://www.annualreviews.org/content/journals/10.1146/annurev.psych.121208.131628 doi: 10.1146/annurev.psych.121208.131628

Davis, I., Carlson, R., Dunham, Y., & Jara-Ettinger, J. (2023). Identifying social partners through indirect prosociality: A computational account. *Cognition*, *240*, 105580.

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12.

Delton, A. W., Jaeggi, A. V., Lim, J., Sznycer, D., Gurven, M., Robertson, T. E., ... Tooby, J. (2023). Cognitive foundations for helping and harming others: Making welfare tradeoffs in industrialized and small-scale societies. *Evolution and Human Behavior*, *44*(5), 485–501.

Delton, A. W., & Robertson, T. E. (2016). How the mind makes welfare tradeoffs: evolution, computation, and emotion. *Current Opinion in Psychology*, *7*, 12-16. doi: 10.1016/j.copsyc.2015.06.006

Eisenbruch, A. B., & Krasnow, M. M. (2022). Why warmth matters more than competence: A new evolutionary approach. *Perspectives on Psychological Science*, *17*(6), 1604–1623.

Eisenhardt, K. M., & Zbaracki, M. J. (1992). Strategic decision making. *Strategic Management Journal*, *13*(S2), 17–37. doi: 10.1002/smj.4250130904

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.

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.

Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, *32*(10), 1402–1413. doi: 10.1177/0146167206291006

Horner, M., Quillien, T., & Moore, A. (2025). Exploring the intuitive theory of empathy. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*. Retrieved from https://escholarship.org/uc/item/9gw0888x

Houlihan, S. D., Ong, D., Cusimano, M., & Saxe, R. (2022). Reasoning about the antecedents of emotions: Bayesian causal inference over an intuitive theory of mind. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*. Retrieved from https://escholarship.org/uc/item/7sn3w3n2

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

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

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. doi: 10.1016/j.cognition.2017.06.017

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 39).

Kuznetsova, A., Brockhoff, P. B., & B.Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Lim, J. (2012). Welfare tradeoff ratios and emotions: Psychological foundations of human reciprocity. *Welfare tradeoff ratios and emotions: psychological foundations of human reciprocity*.

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

Mercier, M., Morin, O., Mercier, H., & Quillien, T. (2026). Who knows what? Bayesian competence inference guides knowledge attribution and information search. *Cognition*, *273*, 106533.

Ong, D. C., Zaki, J., & Goodman, N. D. (2018). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in Cognitive Science*, *11*(2), 338–357. doi: 10.1111/tops.12371

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*.

Powell, L. J. (2022). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*, *17*(5), 1215–1233.

Prelec, D. (1998). The Probability Weighting Function. *Econometrica*, *66*(3), 497–527. Retrieved 2025-12-03, from `https://www.jstor.org/stable/2998573` doi: 10.2307/2998573

Qi, W., Vul, E., & Powell, L. J. (2025). An accurate and efficient measure of welfare tradeoff ratios. *PLOS One*, *20*(5). doi: 10.1371/journal.pone.0322410

Quillien, T. (2023). Rational information search in welfare-tradeoff cognition. *Cognition*, *231*, 105317.

Quillien, T., Tooby, J., & Cosmides, L. (2023). Rational inferences about social valuation. *Cognition*, *239*, 105566. doi: 10.1016/j.cognition.2023.105566

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.

Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*(1394), 427–431. doi: 10.1098/rspb.1998.0312

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

Stengård, E., Juslin, P., Hahn, U., & Van den Berg, R. (2022). On the generality and cognitive basis of base-rate neglect. *Cognition*, *226*, 105160.

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), 72–81.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, *22*.

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational inference of beliefs and desires from emotional expressions. *Cognitive Science*, *42*(3), 850–884.