Exploring the Intuitive Theory of Empathy

Madeleine Horner (m.horner-1@ed.ac.uk),
Tadeg Quillien (tadeg.quillien@ed.ac.uk), **& Adam Moore (amoore23@ed.ac.uk)**
University of Edinburgh, Edinburgh, Scotland

Abstract

Empathy is an emotion that plays a key role in emotional understanding and perspective-taking, and has been identified as a strong motivator for prosocial behavior. We explore people's intuitive theory of empathy, focusing more specifically on the role that the concept of empathy plays in people's causal model of prosocial behavior. We suggest that people implicitly think of empathy as indexing the weight that the actor puts on the welfare of the recipient when deciding whether to help. We test this proposal by asking participants (N=150) to read a series of vignettes in which an actor has the opportunity to help a recipient in need. We find that participants have a robust expectation that actors who feel empathy for the recipient are more likely to help. Furthermore, participants seem to expect that actors who feel empathy are more sensitive to the potential benefits of an action when deciding whether to help. We also test if people can 'invert' this intuitive theory to make inferences about an actor's empathy, given their observable behavior. We find only weak evidence that they can do so, although this might be due to limitations in our experimental design. Overall, our work is a first step toward elucidating the computational principles underlying laypeople's conception of empathy.

Keywords: Empathy; Bayesian modeling; Prosocial behavior; Inference; Emotion

Introduction

Empathy is, broadly, the ability to relate and react to the emotional states of others (Omdahl, 2014; Zaki & Ochsner, 2012). For example, seeing someone in distress can in turn make us feel distressed. The literature on empathy is extensive, but has yet to reach a formal consensus on what empathy is, how it functions, or how people intuitively understand empathy and its import. While cognitive neuroscience has taken a physiological approach focusing on identifying neural correlates of empathy (Weisz & Zaki, 2018), cognitive science has focused on the creation of computational models of empathy, albeit limited by the complexity of conceptualizing empathy (Yalçın & DiPaola, 2020). In contrast, social psychology takes a step back from the individual to focus on how empathy influences interpersonal and group dynamics (Bruneau, Cikara, & Saxe, 2017).

Many features of empathy are highly conserved across species, suggesting that the emotion plays an important evolutionary function (Decety, Bartal, Uzefovsky, & Knafo-Noam, 2016). Evolutionary psychologists have suggested that empathy is a naturally evolved cognitive mechanism (De Waal, 2008) playing a key role in emotional understanding and perspective taking (Yalçın & DiPaola, 2020). An important function of empathy might be to guide prosocial behavior (Goetz, Keltner, & Simon-Thomas, 2010). Altruism can have fitness benefits for a variety of factors, such as kin selection and reciprocal exchange, leading to the evolution of cognitive adaptations for helping others (Delton et al., 2023). By highlighting when someone else is in need, empathy plays a role in steering our helping efforts to cases where help is potentially most beneficial (Sznycer, Delton, Robertson, Cosmides, & Tooby, 2019; Goetz et al., 2010; Delton, Petersen, DeScioli, & Robertson, 2018). Accordingly, empathy has been linked to helping behavior across many empirical studies (Batson, Duncan, Ackerman, Buckley, & Birch, 1981; Coke, Batson, & McDavis, 1978; Weisz, Ong, Carlson, & Zaki, 2021; Decety et al., 2016).

Because of the link between empathy and prosocial behavior, some have suggested that interventions making people more empathetic might be highly socially beneficial (Zaki, 2020). However, other researchers have argued that empathy might lead people to focus their helping efforts in parochial or self-interested ways that might conflict with broader social welfare (Delton et al., 2018; Bruneau et al., 2017; Bloom, 2017).

Given the wide interest in the emotion, we believe it is important to also understand laypeople's intuitive understanding of empathy. Our work is a contribution to a nascent literature in cognitive science on people's 'intuitive theories' of emotions (Wu, Baker, Tenenbaum, & Schulz, 2018; Ong, Zaki, & Goodman, 2015, 2019; Houlihan, Kleiman-Weiner, Hewitt, Tenenbaum, & Saxe, 2023; Smith-Flores, Bonamy, & Powell, 2023). Within this framework, people's lay emotion concepts are understood with respect to the functional role they play within a causal theory of how the minds of other people work (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Houlihan et al., 2023; Smith-Flores & Powell, 2023). In this paper, we are interested in the role that the intuitive concept of empa-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

thy plays in the causal model that people use to explain and predict prosocial behavior.

Utility calculus and welfare-tradeoff ratios

When people predict and explain the behavior of others, they implicitly assume that agents make decisions in an approximately rational manner based on the expected utility or value they place on an action (Lucas et al., 2014; Jara-Ettinger et al., 2016). Within this intuitive causal model of behavior, prosocial behavior can be modeled by assuming that the actor decides to act based in part on the potential benefits of an action to a recipient (Quillien, 2023; Powell, 2022; Eisenbruch & Krasnow, 2022; Qi & Vul, 2022). Formally, the actor assigns a weight λ to the recipient's welfare, and seeks to maximize the utility function:

$$U_{\text{action}} = \lambda B_{\text{Recipient}} - C_{\text{Actor}} \tag{1}$$

where $B_{\text{Recipient}}$ is the potential benefit to the recipient, and C_{Actor} the cost paid by the actor for helping (Quillien, Tooby, & Cosmides, 2023). Parameter λ is a welfare-tradeoff ratio (WTR), that represents how much the actor values the recipient's welfare (Sell et al., 2017; Hall, Kahn, Skoog, & Oberg, 2021; Delton et al., 2023).

Our proposal is that the functional role of the concept of empathy in people's causal model of prosocial behavior can be described computationally in terms of the welfare-tradeoff ratio. That is, we suggest that people (implicitly) think that empathy indexes an actor's welfare-tradeoff ratio toward the potential recipient. People think that (all else being equal), someone who feels empathy toward a potential recipient of help puts a higher value on the recipient's welfare.

Combined with the causal model in Equation 1, this hypothesis generates the following predictions. First, people should expect that actors who feel empathy are more likely to help (since higher WTRs result in higher probability of helping). Second, people should expect that actors who feel empathy are more sensitive to the potential benefits of an action when deciding whether to help. Third, people should also expect that actors are more likely to help when the benefits of helping are high and its costs are low.

Bayesian inference of empathy

As a corollary, we also look at whether people can 'invert' this intuitive theory to infer if someone is feeling empathy based on their observable prosocial actions. While prosocial actions are often associated with empathy, empathy is an internal state which is not observable. Observers must infer empathy indirectly through observable behavior. Previous research suggests that people can make inferences about the emotions of others, often on the basis of little information (Wu et al., 2018; Houlihan et al., 2023; Smith-Flores & Powell, 2023). We suggest that there is an evolutionary benefit to correctly identifying whether someone feels empathy in a given circumstance, given that this information can help predict whether that person will help in the future (Mafessoni & Lachmann, 2019). We investigate if people make inferences about empathy that are consistent with Bayesian reasoning, as described in Equation 2:

$$P(\text{Empathy} \mid \text{Action}, \text{Situation}) = \frac{P(\text{Action} \mid \text{Empathy}, \text{Situation}) \cdot P(\text{Empathy} \mid \text{Situation})}{P(\text{Action} \mid \text{Situation})} \tag{2}$$

Bayes' rule provides a structured framework to look at how people incorporate their prior beliefs about empathy with observable prosocial actions to update their beliefs about the internal state of an actor when doing an action. People's inferences may provide insights into the strength of empathy as a social signal as well the strength of prosocial behavior in signaling internal states.

Current Study

Here we test our hypotheses by asking participants to read about situations where a potential *recipient* is in need, and an *actor* has the opportunity to help the recipient. For various possible helpful actions, we ask participants how costly it would be to perform that action, and how beneficial it would be to the recipient. We also ask them to assess the probability that the actor will help. Our framework predicts that participants should have the following expectations:

- Actors who feel empathy for the recipient are more likely to help,
- Actors are more likely to help when i) the action cost is low and ii) the action benefit is high,
- The benefit of an action should motivate prosocial behavior more strongly in actors who feel empathy.

We also examine whether people make inferences about an actor's level of empathy given their prosocial actions. We do so by asking participants about the probability that an actor feels empathy for the recipient, given what the actor did. Following Wu et al. (2018), we compare these posterior probability judgments to the predictions of a model that combines participants' other probability judgments using Bayes' rule.

Methods

Ethics and Open Science

We preregistered this study's procedure, hypotheses, and analysis plan on the Open Science Framework (see `https://osf.io/rmh8b`). This study also received ethical approval (Approval number: 106-2425/6) and follows all British Psychological Society guidelines. Data and R

code for modeling and analysis are available on the OSF at https://osf.io/r7pmj.

Participants

Participants were recruited using Prolific. We collected data from two groups of participants. In Group A, 50 participants (29 Female, Mean Age = 40.28) were recruited to provide cost/benefit ratings and judgments of prior probability. In Group B, 100 participants were recruited to provide judgments of posterior probability, action likelihood, and marginal action probability (52 Female, Mean age = 39.11). Both groups of participants were compensated £1.05. All participants signed consent forms prior to participation.

Materials

Participants read short vignettes about two fictitious people, John and his friend Bob. The vignettes described a situation that John found himself in and a corresponding prosocial action that his friend Bob could take to help John. We used a total of 10 different situations (e.g. John's dog was hit by a car) and two possible corresponding actions per situation, one high cost (e.g. John's friend Bob took off from work to watch the dog) and one low cost (e.g. John's friend Bob donated $20 to a GoFundMe to help cover vet bills). Participants were told to treat each situation $\times$ action combination independently, as if they featured different characters each time. All materials for this study are available at https://osf.io/r7pmj.

We collected data from two separate groups of participants. Participants in Group A rated the prior probability that Bob would feel empathy for John in that particular situation (they made this rating before reading about the corresponding actions). They also rated the costs and benefits of the actions Bob could take to help John. Participants in Group B provided ratings of posterior probability, action likelihood, and marginal action probability (see details below).

Procedure

Participants in Group A read the 10 vignettes, presented in random order. For each vignette, they were first asked to provide a rating of the prior $Pr(\text{Empathy}|\text{Situation})$. That is, they were asked how likely it is that a person would feel empathy for their friend in this situation, on a 1-100 scale. Participants were then asked to consider two possible actions (presented in random order) that someone could take to help John in this situation. For each action, participants gave ratings of inconvenience (How inconvenient do you think it would be to perform this action?), effort (How much effort do you think this action requires?), emotional benefit (How emotionally beneficial do you think this action would be to the recipient?), and practical benefit (How practically beneficial do you think this action would be to the recipient?). Each of these ratings were measured on a 7pt. Likert scale ranging from 'none at all' to 'a great deal'.

Participants in Group B also read the same 10 vignettes. However, in contrast to Group A, each participant was only shown one action per situation—either the high-cost or the low-cost action. Participants were shown five high cost actions and 5 low cost actions. For each action they gave a rating of the marginal probability $Pr(\text{Action}|\text{Situation})$, the likelihood $Pr(\text{Action}|\text{Empathy, Situation})$, and the posterior $Pr(\text{Empathy}|\text{Action, Situation})$, all on a 1-100 scale. Concretely, they were asked to rate the probability that someone in Bob's situation would help ('Out of 100 people in Bob's situation, how many would do what he did if they had the opportunity?'; marginal probability), the same probability, conditional on having empathy ('Out of 100 people in Bob's situation, how many would do what he did if they had the opportunity and felt empathy for their friend?'; likelihood), and the probability that Bob feels empathy given what he did ('How likely is it that Bob feels empathy for John?'; posterior).

Results

Analyses were performed with `brms` (version 2.22.0) in `R` (version 4.3.2) (Bürkner, 2017). Pre-registered Bayesian multiple regression models are reported with standardized β coefficients. Model comparison, where appropriate, was based on leave-one-out-cross-validated log-pointwise predictive density (LOO-ELPD) estimates and Bayes Factor comparisons. Parameter estimation is based on the mean of the posterior distributions and 95% highest density intervals. Theoretically motivated predictions were operationalized as priors on standardized β coefficients, which are available on OSF (see https://osf.io/rmh8b).

Our analyses focused on two broad questions: (1) what factors predict how probable participants judge a particular helping behavior to be and (2) can people infer empathy in a target based on prosocial actions.

Predictive Power of Empathy

We find that participants have a robust expectation that actors who feel empathy are more likely to help. Figure 1 shows that the effect consistently appears in each of our ten vignettes (each bar is above 1). Figure 2 gives a sense of the magnitude of the effect. Telling participants that the actor feels empathy (green line) moderately increased their estimate that the actor would help, relative to no information (blue line). While we did not ask participants about an actor who feels no empathy, we can 're-construct' the corresponding probability estimates by computing the estimates of $Pr(\text{Action}|\neg\text{Empathy}, \text{Situation})$ that were consistent with the other probability judgments given by participants. These re-constructed judgments (red line) are much

lower than judgments for $Pr(\text{Action}|\text{Empathy}, \text{Situation})$, see Figure 2.

To investigate question (1) more exhaustively, we used the cost and benefit ratings from Group A participants, along with the presence/absence of empathy information in prompts for Group B participants' judgments, as predictors for Group B participants' action probability ratings. As outlined above, participants in Group B provided two ratings of the probability that Bob would perform an action in a given situation: the likelihood $Pr(\text{Action}|\text{Empathy, Situation})$ and the marginal probability $Pr(\text{Action}|\text{Situation})$. Hence, we define an 'Empathy' variable that has value 1 for questions where participants were asked about the likelihood $Pr(\text{Action}|\text{Empathy, Situation})$, and 0 for questions where participants were asked about the marginal probability $Pr(\text{Action}|\text{Situation})$. This allows us to examine whether participants think that a person is more likely to perform a prosocial action if that person feels empathy for the recipient (relative to the base case where information about empathy is absent).

We ran a Bayesian multi-level model predicting action probability with fixed effects of empathy, cost, and benefit, and participant-level random effects.[1] Participants consistently expected empathy to increase the probability that an actor would perform an action ($\beta = 0.35$, 95% CI $[0.29, 0.41]$, $SE = 0.03$) and that the more costly an action, the lower the probability that people thought an actor would perform it ($\beta = -0.26$, 95% CI $[-0.34, -0.18]$, $SE = 0.04$). However, contrary to our predictions, the benefit of an action was not related to participants' probability judgments regarding action performance ($\beta = -0.03$, 95% CI $[-0.13, 0.06]$, $SE = 0.05$).

Importantly, we found a positive interaction between empathy and benefit ($\beta = 0.07$, 95% CI $[0.01, 0.12]$, $SE = 0.03$): Perceived benefit had a more positive weight in people's predictions when they were told the actor felt empathy for the recipient. This effect was robust to prior specification with an 87% overlap in the 95% highest density intervals between models using informed and weakly informative priors (see `https://osf.io/r7pmj`). This finding provides some support for the proposal that empathy indexes the welfare trade-off ratio of the actor for the recipient. This is because, in the generative model in Equation 1, the welfare-tradeoff ratio λ indexes the weight the actor places on the potential benefit to the recipient. This equation makes apparent that if people think that empathy indexes the welfare-tradeoff ratio, empathy should modulate the importance given to perceived benefit in predicting the decision to help.

[1] Deviating from our pre-registration, we dropped situation-level random effects from the pre-registered models. The pre-registered model fits 10 intercepts for each situation but there is only one data point per participant for each situation, resulting in significant data shrinkage. We also used the means of cost and benefit, which was computed by creating a composite score for cost (e.g. effort and inconvenience) and benefit (e.g. practical and emotional benefit) and then finding the mean cost and benefit rating per action by situation and level, either high or low.

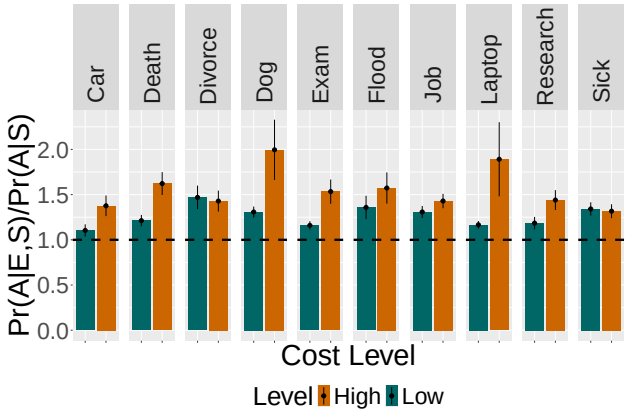

Figure 1: Ratio between judged action probability when the actor feels empathy and judged action probability when no information about empathy is provided, across situations and cost levels. Color indexes our cost manipulation: teal bars show the participants' judgments about low cost actions, and orange bars show the participants' judgments about high cost actions. Each panel represents a different situation. Error bars represent standard errors.

In an exploratory analysis, we find that people expect empathy to increase the probability of helping more for high-cost than for low-cost actions, see Figure 1. Specifically, we ran a multi-level model predicting the ratio $Pr(\text{Action} \mid \text{Empathy}, \text{Situation})/Pr(\text{Action} \mid \text{Situation})$ as a function of our cost manipulation, with participant- and situation-level random intercepts. We find that this ratio is higher in the high-cost than in the low-cost condition, $\beta = 0.23$, 95% CI $[0.15, 0.32]$, $SE = 0.04$.

In sum, our main findings in this section are the following. Costly actions were judged unlikely to be undertaken, providing some evidence for people making judgments on the basis of utility maximization. Empathy was consistently judged to be a meaningful predictor of prosocial action. Information that the actor felt empathy increased the weight of perceived benefit in people's judgment of the probability that the actor would help.

Bayesian Inference of Empathy

Participants seem to have a robust expectation that those who feel empathy are more likely to help. Question (2) is whether people can 'invert' the generative model in their intuitive theory of empathy to infer if someone feels empathy based on their prosocial actions. Our target in these analyses is participants' posterior judgments, i.e. their ratings of $Pr(\text{Empathy}|\text{Action}, \text{Situation})$. We first constructed a main computational model that assumes that par-

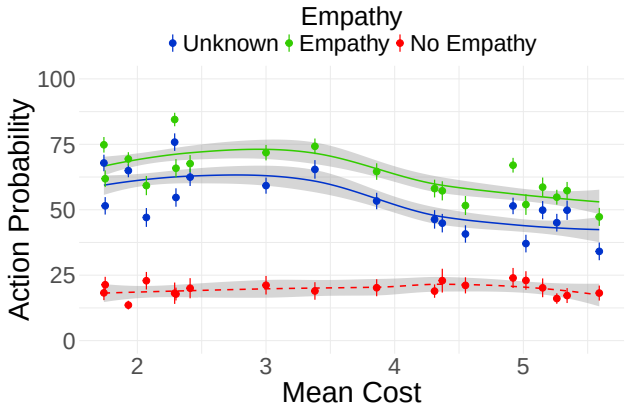

Figure 2: Average judged action probability as a function of the mean perceived cost of the action, and information about the actor's empathy toward the recipient. The blue line represents the judged marginal probability of an action $Pr(\text{Action}|\text{Situation})$ while the green line represents the judged action likelihood $Pr(\text{Action}|\text{Empathy}, \text{Situation})$. The red line represents the (reconstructed) judged probability of an action if the actor felt no empathy $Pr(\text{Action}|\neg\text{Empathy}, \text{Situation})$. We imputed these values on the basis of the other judgments we collected, as $Pr(\text{Action}|\neg\text{Empathy}, \text{Situation}) = (Pr(\text{Action} \mid \text{Situation}) - \Pr(\text{Action} \mid \text{Empathy}, \text{Situation}) \cdot \Pr(\text{Empathy})/(Pr(\neg\text{Empathy})$.

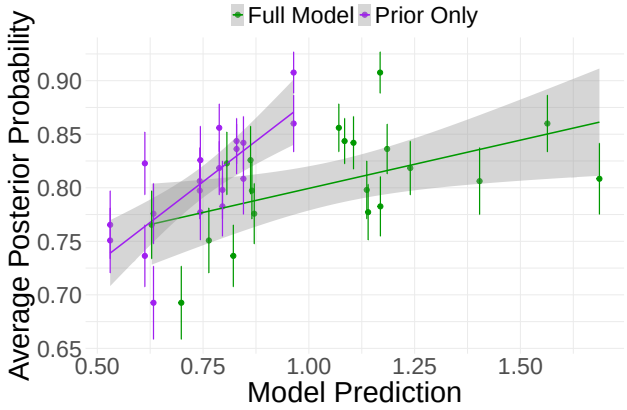

Figure 3: Average human posterior judgments, as a function of the predictions made by the full model (green), and the prior-only lesion model (purple). Each dot represents an action $\times$ situation combination. Error bars represent standard errors.

ticipants compute this posterior according to Bayes' rule:

$$P(\text{Empathy} \mid \text{Action}, \text{Situation}) = \frac{P(\text{Action} \mid \text{Empathy}, \text{Situation}) \cdot P(\text{Empathy} \mid \text{Situation})}{P(\text{Action} \mid \text{Situation})} \tag{3}$$

For each action $\times$ situation combination we generated the prediction of this model by first averaging participants' ratings for each of the relevant probability judgments (i.e. $Pr(\text{Action}|\text{Empathy}, \text{Situation})$, $Pr(\text{Action}|\text{Situation})$ and $Pr(\text{Empathy}|\text{Situation})$)[2], and then combining these average ratings as in the right-hand-side of Equation 3.

We also tested two 'lesioned' versions of this model. The 'prior-only' model assumes that posteriors only track the corresponding prior:

$$P(\text{Empathy} \mid \text{Action}, \text{Situation}) = P(\text{Empathy} \mid \text{Situation}) \tag{4}$$

While the 'ratio' model assumes that people track the ratio of the likelihood to the marginal probability and ignore the prior:

$$P(\text{Empathy} \mid \text{Action}, \text{Situation}) = \frac{P(\text{Action} \mid \text{Empathy}, \text{Situation})}{P(\text{Action} \mid \text{Situation})} \tag{5}$$

We report the fit of each computational model to the full dataset using a Bayesian multi-level regression[3], as well as the Pearson's correlation between model prediction and the average item-level human judgment. For purposes of model comparison we also assess model fit using WAIC and ELPD-LOOIC (computed from the Bayesian multi-level regression), with lower values indicating better model fit.

The predictions of the full model were moderately correlated with human judgments, item-level correlation: $r = 0.40$, 95% CI $[0.06, 0.71]$, full dataset: $\beta = 0.12$, 95% CI $[0.08, 0.17]$, $SE = 0.02$, $WAIC = 2407.9$, $LOO = 2409.3$. Contrary to our predictions, the prior-only model had the best fit to the data, item-level correlation: $r = 0.66$, 95% CI $[0.38, 0.85]$, full dataset: $\beta = 0.18$, 95% CI $[0.12, 0.23]$, $SE = 0.02$, $WAIC = 2374.0$, $LOO = 2375.5$. The ratio model had the weakest fit to the data, item-level: $r = 0.20$, 95% CI $[-0.18, 0.56]$, full dataset: $\beta = 0.07$, 95% CI $[0.03, 0.12]$, $SE = 0.03$, $WAIC = 2425.8$, $LOO = 2427.5$.

These results suggest that participants relied mainly on their prior beliefs about the expected level of empathy given the situation when judging if someone is feeling empathy, see Figure 3. We find at best weak evidence that participants also used what the actor did to infer whether they felt empathy. We discuss possible interpretations of

[2]Participants provided very low ratings for the marginal probability $Pr(\text{Action}|\text{Situation})$ for most items, which resulted in 'impossible' probability values (larger than 1) for these items after combining the probability ratings, see Figure 3. Incoherence between probability judgments is not altogether unexpected in human probabilistic reasoning (Zhu, Sanborn, & Chater, 2020).

[3]Specified as the posterior judgments given by participants predicted by the fixed effect of the model prediction, either the full model or one of the two lesioned models, and participant-level random effects

these results in the Discussion.

Discussion

Empathy is an emotion that plays a key role in social behavior and has attracted considerable attention from researchers. Past research has, for example, suggested that empathy is a strong motivator for prosocial behavior (Batson et al., 1981; Coke et al., 1978; Weisz et al., 2021; McCauley, McAuliffe, & McCullough, 2024). In this paper, we start to explore people's intuitive conception of empathy. We take a computational approach, exploring the functional role that the concept of empathy plays in the causal model that people use to predict and explain behavior. In particular we are interested in the context of predicting prosocial behavior.

We provide empirical evidence that people systematically expect feeling empathy to be associated with more prosocial behavior, suggesting that people think of empathy as indexing an actor's welfare-tradeoff ratio (WTR) toward the recipient (Quillien et al., 2023). The WTR perspective makes a further prediction for which we also find some evidence: People expect that empathy increases the actor's sensitivity to an action's potential benefits when deciding whether to help. That is, people do not simply have a blanket assumption that empathy is linked to prosocial behavior, but also have more fine-grained expectations.

Our work is inspired by a substantial body of research elucidating the causal models and planning algorithms that support social cognition (Crockett, 2013; Cushman, 2013; Sloman, Fernbach, & Ewing, 2009, 2012; Jara-Ettinger et al., 2016; Quillien & German, 2021). In particular, we contribute to an ongoing effort to map the functional role of emotion concepts in people's intuitive psychology (Ong et al., 2015; Houlihan et al., 2023). Our results are broadly consistent with the idea that common-sense psychological reasoning is supported by a 'naive utility calculus', whereby people anticipate other agents to act based on the expected utility they assign to actions (Lucas et al., 2014; Jara-Ettinger et al., 2016).

Although participants in our study consistently associate empathy with prosocial behavior, we do not find strong evidence that they use this expectation to infer empathy from observable behavior. Specifically, a model that predicted participants' inferences on the basis of situation-specific priors provided a better account than a Bayesian model that also incorporated the relevant action probabilities.

One possible explanation for this result is that some of the assumptions behind our modeling approach might not hold. In particular, we collected judgments of prior probability from one group of participants, and assumed that these judgments would reflect the priors of participants who had to judge the posterior probability that an actor feels empathy on the basis of what the actor did. But a description of an action might also contain cues about the situation (for example the fact that someone is contemplating a very costly action might hint at the severity of the recipient's need) and change participants' priors.

Participants may also be using approximation strategies to make their probability judgments. For example, they might be using simple counterfactual-based inferences that do not result in normative probability estimates. They might also be using some kind of sampling algorithm (Stewart, Chater, & Brown, 2006; Stewart, 2009; Bhui & Gershman, 2018; Bramley, Zhao, Quillien, & Lucas, 2023; Zhu et al., 2020) to evaluate the extent to which empathy is present when actions similar to those described are undertaken. Such sampling is often cue driven, and the stimuli themselves may have prompted participants to sample extreme events as more definitive and informative, biasing the samples from which they derived their priors, likelihoods, and thus posterior inferences (Kvam, Alaukik, Mims, Martemyanova, & Baldwin, 2022; Lieder, Griffiths, & Hsu, 2018) and resulting in inferences that do not follow a normative Bayesian prediction.

We also note that our experiment provided a relatively stringent test of people's capacity to make action-based inferences of empathy: we only asked people to consider situations where the actor did help. As such our study only allows us to assess whether people treat one particular helpful action as more diagnostic about empathy than another helpful action. This resulted in relatively weak variation in the diagnosticity of different observed actions, relative to the variance in participants' situation-specific priors.

Limitations and Future Directions

While the diversity of situations presented in this study is useful in allowing us to probe the robustness of the effects found, this comes at the cost of tight experimental control of the cost and benefit of the associated actions. Future work could use more controlled material, for example to systematically manipulate cost and benefit in an orthogonal way. It will also be fruitful to study whether an actor's refusal to help impacts an observer's inference about the actor's empathy.

The present work is a first step within a potentially large research program. A full investigation of people's lay conceptions should also explore their beliefs about the *antecedents* of empathy: what do people think gives rise to the emotion? For example, how do people conceptualize the notion of 'need', and do they expect empathy to be biased toward close partners? Answering these and other questions will be key to mapping out the intuitive theory of empathy.

References

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T.,

& Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of personality and Social Psychology*, *40*(2), 290.

Bhui, R., & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, *125*(6), 985-1001.

Bloom, P. (2017). *Against empathy: The case for rational compassion*. Random House.

Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (2023). Local search and the evolution of world models. *Topics in Cognitive Science*.

Bruneau, E. G., Cikara, M., & Saxe, R. (2017). Parochial Empathy Predicts Reduced Altruism and the Endorsement of Passive Harm. *Social Psychological and Personality Science*, *8*(8), 934.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Coke, J. S., Batson, C. D., & McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of personality and social psychology*, *36*(7), 752.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363-366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273-292.

Decety, J., Bartal, I. B.-A., Uzefovsky, F., & Knafo-Noam, A. (2016). Empathy as a driver of prosocial behaviour: Highly conserved neurobehavioural mechanisms across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150077.

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., Petersen, M. B., DeScioli, P., & Robertson, T. E. (2018). Need, compassion, and support for social welfare. *Political Psychology*, *39*(4), 907–924.

De Waal, F. B. (2008). Putting the Altruism Back into Altruism: The Evolution of Empathy. *Annual Review of Psychology*, *59*(1), 279–300.

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.

Goetz, J. L., Keltner, D., & Simon-Thomas, E. (2010). Compassion: an evolutionary analysis and empirical review. *Psychological bulletin*, *136*(3), 351.

Hall, J., Kahn, D. T., Skoog, E., & Oberg, M. (2021). War exposure, altruism, and the recalibration of welfare tradeoffs towards threatening social categories. *Journal of Experimental Social Psychology*, *94*.

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 naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

Kvam, P. D., Alaukik, A., Mims, C. E., Martemyanova, A., & Baldwin, M. (2022). Rational inference strategies and the genesis of polarization and extremism. *Scientific reports*, *12*(1), 7344.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*(1), 1.

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.

Mafessoni, F., & Lachmann, M. (2019). The complexity of understanding others as the evolutionary origin of empathy and emotional contagion. *Scientific Reports*, *9*(1), 5794.

McCauley, T. G., McAuliffe, W. H., & McCullough, M. E. (2024). Does empathy promote helping by activating altruistic motivation or concern about social evaluation? a direct replication of fultz et al.(1986). *Emotion*.

Omdahl, B. L. (2014). *Cognitive appraisal, emotion, and empathy*. Psychology Press.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

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.

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

Qi, W., & Vul, E. (2022). The evolution of theory of mind on welfare tradeoff ratios. *Evolution and Human Behavior*, *43*(5), 381–393.

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

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

Quillien, T., Tooby, J., & Cosmides, L. (2023). Rational inferences about social valuation. *Cognition*, *239*, 105566.

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.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. *Psychology of learning and motivation*, *50*, 1–26.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2012). A causal model of intentionality judgment. *Mind & Language*, *27*(2), 154–180.

Smith-Flores, A. S., Bonamy, G. J., & Powell, L. J. (2023). Children's reasoning about empathy and social relationships. *Open Mind*, *7*, 837–854.

Smith-Flores, A. S., & Powell, L. J. (2023). Joint reasoning about social affiliation and emotion. *Nature Reviews Psychology*, *2*(6), 374–383.

Stewart, N. (2009). Eps prize lecture: Decision by sampling: The role of decision environment in risky choice. *Quarterly Journal of Experimental Psychology*, *62*(6), 1041-1062.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1-26.

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.

Weisz, E., Ong, D. C., Carlson, R. W., & Zaki, J. (2021, August). Building empathy through motivation-based interventions. *Emotion*, *21*(5), 990–999.

Weisz, E., & Zaki, J. (2018). Motivated empathy: A social neuroscience perspective. *Current Opinion in Psychology*, *24*, 67–71.

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.

Yalçın, Ö. N., & DiPaola, S. (2020). Modeling empathy: Building a link between affective and cognitive processes. *Artificial Intelligence Review*, *53*(4), 2983–3006.

Zaki, J. (2020). *The war for kindness: Building empathy in a fractured world*. Crown.

Zaki, J., & Ochsner, K. N. (2012, May). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience*, *15*(5), 675–680.

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.