Who went fishing? Inferences from social evaluations

Zachary J. Davis (zach.davis@stanford.edu)
Stanford University, Department of Psychology

Kelsey Allen (krallen@mit.edu)
Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences

Tobias Gerstenberg (gerstenberg@stanford.edu)
Stanford University, Department of Psychology

Abstract

Humans have a remarkable ability to go beyond the observable. From seeing the current state of our shared kitchen, we can infer what happened and who did it. Prior work has shown how the physical state of the world licenses inferences about the causal history of events, and the agents that participated in these events. Here, we investigate a previously unstudied source of evidence about what happened: social evaluations. In our experiment, we present situations in which a group failed to optimally coordinate their actions. Participants learn how much each agent was blamed for the outcome, and their task is to make inferences about the situation, the agents' actions, as well as the agents' capabilities. We develop a computational model that accurately captures participants' inferences. The model assumes that people blame others by considering what they should have done, and what causal role their action played. By inverting this generative model of blame, people can figure out what happened.

Keywords: blame attribution; counterfactual reasoning; social cognition; Bayesian inference; computational modeling.

Introduction

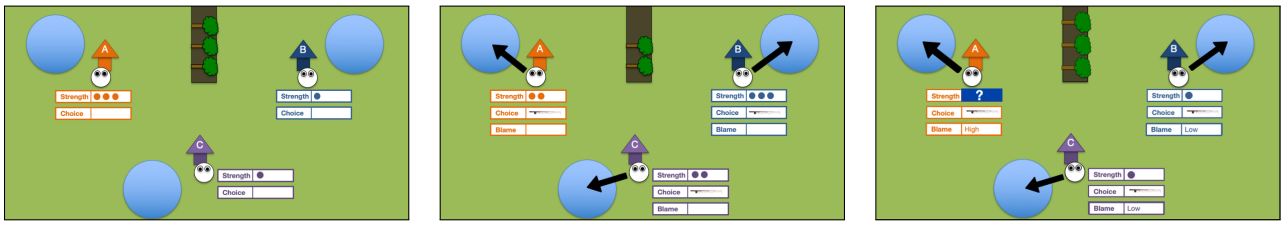
A remarkable aspect of human intelligence is the ability to draw sophisticated inferences that go beyond what can be perceived directly. For example, from observing the current state of the physical world, people can infer what must have happened in the past (Smith & Vul, 2014; Gerstenberg, Siegel, & Tenenbaum, 2018; Kirfel et al., 2020). These inferences about the past include physical events (e.g. the wind must have pushed the window open), as well as events involving other people (e.g. my roommate must have been hungry at night and left the fridge open). Research has shown how people can use physical evidence to infer what actions agents took (Schachner & Kim, 2018), what goals they had (Lopez-Brau et al., 2020), and what their knowledge states were (Pelz et al., 2020). To infer what happened, people also naturally draw on psychological evidence, such as the emotional expressions of others (Weiner, 1985; Wu et al., 2021). For example, a sad and disappointed sports fan reveals who won the game. Here, we investigate a source of information about what happened that hasn't been explored: social evaluations.

Humans are evaluative creatures and social evaluations, such as attributions of responsibility and blame, form an important part of our everyday lives (Malle, 2021; Alicke et al., 2015; Lagnado et al., 2013). Social evaluations provide a rich source for inferences about what happened, because the way in which we hold each other accountable for our actions follows systematic patterns. If people's intuitions about how blame should be attributed are generally shared, then knowing someone else's blame judgment provides diagnostic information about what happened. For example, hearing that a soccer defender was blamed for the team's loss leads to a different picture of what happened than hearing that the striker was blamed. The systematic factors that determine how blame is allocated include normative considerations about how a person should have acted, as well as a consideration of what causal role the person's action played in bringing about the outcome (see Malle, 2021; Gerstenberg, Ullman, et al., 2018).

Here, we focus on people's inferences from social evaluations for a class of situations in which a group failed to optimally coordinate their actions. This may happen, for example, when a sports team fails to coordinate on their defense, or when a political party fails to pass some legislation. What can people infer from learning how much each individual was blamed for the suboptimal outcome? To model this kind of situation, participants in our experiment were introduced to a village of fishermen (see Figure 1). The fishermen each fish in their own lake, sell their catch, and evenly split their earnings. The fishermen can only sell their fish if the road to the village is not blocked by trees. A fisherman's strength determines how many sacks of fish they can catch if they decide to fish, or how many trees they can remove from the road if they decide to clear the trees. Going fishing or clearing the trees takes all day, so each fisherman has to decide what to do. Because the fishermen live far away from one another, they cannot communicate to coordinate their actions. Each fisherman knows how strong everyone is, and how many trees are blocking the road.

Consider the situation shown in Figure 1a. Fisherman A has strength 3, and fishermen B and C both have strength 1. Three trees are blocking the road. What should each fisherman do in this case? The best possible outcome they can achieve is to sell two sacks of fish. For that to happen, fishermen B and C should go fishing, and fisherman A should go and clear the trees. Here the optimal solution is relatively simple to achieve. However, consider the situation shown in Figure 1b. Here, there are two trees blocking the road and both fishermen A and C have strength two. All three fishermen ended up going fishing. To what extent should each of the fishermen be blamed for the suboptimal outcome?

Our work here builds off of Allen et al. (2015), who in-



(a) **Prediction**: What should each fisherman do?

(b) **Blame**: How much is each fisherman to blame for the suboptimal outcome?

(c) **Inference**: How strong is fisherman A?

Figure 1: The fisherman paradigm. Allen et al. (2015) explored the prediction task (a), and the blame task (b). Here, we focus on the inference task (c): Can people use information about how much each fisherman was blamed to figure out what happened? In this case, the question is whether fisherman A's strength is 1, 2, or 3.

vestigated judgments of blame in the fisherman paradigm. In their Experiment 1, participants saw a number of trials for which they knew each fisherman's strength and how many trees blocked the road, and their task was to say what action each fisherman should take (just like in Figure 1a). In Experiment 2, participants saw situations in which the fishermen failed to achieve the optimal outcome and their task was to judge to what extent fisherman A was to blame for the suboptimal outcome (similar to what's shown in Figure 1b).

To predict participants' judgments of what each fisherman should do in Experiment 1, Allen et al. developed a recursive reasoning model in which each fisherman best responds to what they think the other fishermen will do (explained in more detail below). In Experiment 2, Allen et al. showed that participants' blame judgments were sensitive both to what a fisherman should have done, as well as to the causal role that their action played. Fishermen were blamed more if they would have secured the optimal outcome had they acted differently. So both what a fisherman did and how much it mattered, affected participants' blame judgments. In this paper, we test whether participants can use such blame judgments in the fisherman paradigm to draw inferences. These inferences could be about actions (whether each fishermen went fishing or clearing trees), capacities (how strong each fisherman is), or aspects of the situation (how many trees were blocking the road). Consider the situation shown in Figure 1c. Given that fisherman A was blamed a lot whereas fishermen B and C received little blame, we may be able to infer that fisherman A's strength must have been 3 (or 2) in which case he should have cleared the trees rather than gone fishing.

The rest of the paper is organized as follows. We first lay out the computational modeling framework, discussing both the generative model of blame as well as how that model can be inverted to make inferences about what happened. We then test the model using the fisherman paradigm, asking participants to fill in missing information from judgments of blame. We conclude by discussing limitations of the proposed model, and by sharing future research ideas.

Computational Model

We first outline the generative model of blame which was developed in Allen et al. (2015), and then show how this generative model can be inverted to form a posterior distribution over unknown situational factors.

A generative model of blame

Allen et al.'s (2015) generative model of blame has two components: a *rationality component* which computes whether a person did what they should have done, and a *pivotality component* which computes what causal role a person's action played in bringing about the suboptimal outcome. We describe each component in turn.

Rationality component: What should a person do? In a cooperative situation, agents should act in order to maximize the group's expected reward. A person is blamed to the extent that they failed to act rationally. Assigning blame involves comparing an agent's actions against how they *should have* acted, given their knowledge of the number of fallen trees and strengths of each fisherman. For example, fisherman A in Figure 1a should recognize that only he can clear the trees and take that action himself.

We model rational decision-making as a recursive reasoning process, where each agent acts in response to the others at a level k depth of reasoning (see, e.g. Yoshida et al., 2008). At level $k = 1$, each agent assumes that the other agents act randomly, and chooses the action that maximizes the group's expected reward given the expected actions of the others. At further level k, each agent chooses an action assuming that the other agents have done $k-1$ reasoning. We model the probability of fisherman f choosing action a_f at level k as a softmax distribution with parameter β_r over expected reward outcomes associated with alternative actions:

$$p^k(a_f) = \frac{\exp(\beta_r \hat{r}_k[a_f])}{\sum_{a_f \in \text{actions}} \exp(\beta_r \hat{r}_k[a_f])}, \quad (1)$$

where $\hat{r}_k[a_f]$ is the expected reward for fisherman f to take action a_f based on level k of reasoning. The softmax temperature parameter β_r can range from 0 (ignoring expected rewards and responding randomly) to infinity (deterministically choosing the action with highest expected reward). The

rationality component's predicted blame $\hat{b}$ for fisherman f is their deviation from rational action:

$$\hat{b}_{\text{rationality}} = 1 - p^k(a_f) \tag{2}$$

Pivotality component: How much did a person's action matter? The generative model's rational action component prescribes how someone should act, given a state of uncertainty. However, in retrospect there may be situations where one agent's actions seem reasonable ahead of time but they end up being responsible for the group's failure to sell as many fish as possible. For example, in Figure 1b, it may have been reasonable for A or C to go fishing. Because they both went fishing, however, they are both responsible for the group's failure. The group could have gotten the optimal outcome had one of them decided to go clear the trees instead.

Following Allen et al. (2015), we define the causal responsibility of an agent's action for the outcome by considering how many of the other agents' actions would have needed to be different in order for that agent's action to have been pivotal to the group achieving the optimal outcome (see also Chockler & Halpern, 2004; Lagnado et al., 2013). Formally, the pivotality components' predicted blame $\hat{b}$ for fisherman f is

$$\hat{b}_{\text{pivotality}} = \frac{1}{N_f + 1}, \tag{3}$$

where N_f is the number of other fishermen whose action would have needed to be different in order for the group's success to be counterfactually dependent on the action of fisherman f. Returning to the situation in Figure 1b, the pivotality of fishermen A and C is 1 here because if either of them had acted differently, the group would have achieved the best possible outcome (number of changes to be pivotal $N_f = 0$). If fisherman B had gone to clear the trees in the actual situation, then the pivotality of fishermen A and C would have been 0.5 ($N_f = 1$). Now neither A nor C are pivotal in the actual situation, but would have been pivotal if B's action was changed.

Combining model components Allen et al. (2015) found that participants attributed blame by taking into account both the rationality of an agent's action, as well as what causal role it played. Following them, we model people's judgments of blame for fisherman f as a weighted mixture of the blame values predicted by the rationality and pivotality models

$$\hat{b} = w \cdot \hat{b}_{\text{rationality}} + (1 - w) \cdot \hat{b}_{\text{pivotality}}, \tag{4}$$

with the weighting parameter w controlling how much each component affects the blame judgment.

We will refer to a version of the model in which $w = 1$ as the *rationality model* (because it only considers that component of the model), a version with $w = 0$ as the *pivotality model*, and a version in which $0 < w < 1$ as the *mixture model*.

Inferring what happened by inverting the generative model of blame

The generative model outlined above assigns blame to agents for a given situation. In our setting at least one aspect of the situation is unknown, and the task is to infer the missing pieces of information based on how much blame each agent received together with what was known about the situation (see Figure 1c). For example, in Figure 1c, the only missing piece of information is how strong fisherman A was. There are three possible situations: fisherman A's strength could either be 1, 2, or 3. Computing a probability of a situation s_i, given an assignment of blame to fisherman $p(s_i|b_f)$ involves inverting the generative model of blame using Bayes' rule:

$$p(s_i|b) = \frac{\prod_{f \in \text{fishermen}} \mathcal{L}(b_f|\hat{b}_f) p(\hat{b}_f|s_i) p(s_i)}{\sum_{i \in \text{situations}} \prod_{f \in \text{fishermen}} \mathcal{L}(b_f|\hat{b}_f) p(\hat{b}_f|s_i) p(s_i)}, \tag{5}$$

where s_i is a potential situation, $\hat{b}_f$ is the generative model's predicted blame for fisherman f, and b_f is the blame that the agent actually received. We assume a uniform prior over possible situations $p(s_i)$ but assign 0 probability to situations in which the optimal possible outcome would have been achieved, as participants in our experiment were told that the fishermen were blamed for failing to achieve the best outcome.

The generative model's blame values for each fisherman $p(\hat{b}_f|s_i)$ are deterministic given the parameters of the model. The likelihood function $\mathcal{L}(b_f|\hat{b}_f)$ involves a comparison of the observed blame values b_f against the blame values predicted by the generative model of blame in each situation. In our experiment blame judgments only took 3 values ("low", "medium", or "high"; see Figure 1c). Because the model's predictions are continuous, we converted each qualitative blame judgment into a value (.2, .5, .8, respectively), and computed the likelihood of a model blame value as normally distributed with standard deviation of 0.1 (chosen so that 90% of the probability density function for each qualitative judgment is within its corresponding third of the 0 to 1 scale). This likelihood captures that there is some uncertainty about what continuous value a "low", "medium", or "high" blame judgment maps onto.

Decision noise The final modeling step is to convert a posterior distribution over possible situations to a prediction of the probability with which a participant chooses a response option for each question (such as fisherman A's strength in Figure 1c). When there is more than one unknown feature of a situation (e.g. both fisherman A's and B's actions are unknown), computing the probability of a participant selecting response option o_j for one question (e.g. whether fisherman A went fishing or cleared the trees) involves first marginalizing over other unknowns (fisherman B's action in this case), then softmaxing over the resulting posterior distribution of that factor:

$$p(o_j) = \frac{\exp(\beta_d p(o_j|s_i))}{\sum_{o_j \in \text{options}} \exp(\beta_d p(o_j|s_i))} \tag{6}$$

Experiment

Methods

All materials, data, modeling, and analysis code are available online here: https://github.com/cicl-stanford/inference_from_social_evaluations

Participants 50 participants (24 female, 25 male, 1 prefer not say, age: $M = 41$, $SD = 11$) were gathered on Amazon Mechanical Turk and compensated with $2.75. It took participants 10.2 ($SD = 5$) minutes on average to complete the experiment. 40 additional participants were excluded because they failed the preregistered criterion of passing both of the attention checks (pre-registration: https://osf.io/x37r). In the attention checks, participants were given full information about the scenario and had to correctly say how many fish sacks would be sold.

Procedure & Design Participants were informed of the overall setting of the task and then required to answer comprehension checks that established that they understood that the fishermen shared their earnings equally, two questions establishing how many sacks of fish would be sold in a given situation, and three situations where the number of trees and strengths of all fishermen were known and they were asked "what should each fisherman do, so that together they sell the most fish?". To familiarize them with attributing blame in our setting, participants viewed three situations in which the fishermen failed to achieve the optimal outcome, and judged "how much is each fisherman to blame for the group's failure get the best possible outcome?".

After completing the instructions, participants learned that they would get information about how much each fisherman "was to blame" for the suboptimal outcome, and that their task was to fill in the missing pieces. This wording was designed to be vague in terms of who provided the blame judgment, but make it clear that the fishermen didn't blame each other. For a given trial, participants were presented with images like the one in Figure 1c, with text at the top stating "Try to fill in the missing information". Participants responded using dropdown menus overlaid on the image, options were [1, 2, 3] for a fisherman's strength, ["1 tree", "2 trees", "3 trees"] for fallen trees, and ["trees", "fish"] for a fisherman's actions. 36 trials were presented in random order.

The trials varied what inferences participants were asked to make (trees, choices, and/or strength) and how many of those pieces of information were missing (from one to three). To assess whether participants were sensitive to how much each fisherman was blamed, roughly half of the trials involved cases where the situation was held constant but the amount of blame assigned to each agent varied. For example, in trials 1 and 2 in Figure 2 there was one fisherman with strength 3 and two others with strength 1, and all went fishing. Whereas in trial 1, the weak fishermen received high blame and the strong fisherman received low blame, in trial 2 the pattern was reversed. Similarly, in trial 3 vs. trial 4, the situation was identical (considering the symmetry between fisherman A and B) except for the blame that fishermen A and B received. Both received medium blame in trial 3 versus low and high blame in trial 4.

The rest of the trials were selected so that there was a conflict between how much blame would be assigned according to versions of the model that only consider the rationality component (see Equation 2), or only the pivotality component (see Equation 3). For example, in trial 6 the pivotality model infers that the strength of fisherman B is most likely to have been $S(B) = 2$ because, in this case, his action would have been pivotal (hence the high blame). If fisherman B's strength was 2, then they could have achieved the optimal outcome had he gone for the trees instead of fishing. The rationality model, on the other hand, infers that $S(B) = 3$ is most likely. If $S(B)$ had been 2, it would have been reasonable to go fishing assuming that fisherman C would clear the trees. The reasonableness of B's going fishing in this case would be incompatible with the high blame he received. These trials helped to distinguish between the rationality and pivotality models by reducing the correlation between their predictions which was $r(\text{rationality}, \text{pivotality}) = .57$ across all trials.

Results & Discussion

Figure 2 shows a selection of 8 trials from the experiment. Each trial shows the probability with which participants selected the different response options, together with the model predictions. These results show that participants' inferences are sensitive to the blame information. For example, trials 1 and 2 are identical in terms of the fishermens' strengths and choices (except for shuffling), but differ in how much blame each fisherman received. Whereas participants were most likely to infer that there was one tree in trial 1, they inferred that there must have been three trees in trial 2. The models correctly capture these inferences. Trials 3 and 4 also only differ in the blame information. Even though fisherman C takes the same action and receives the same amount of blame in both settings, participants were able to use information about how much the *other* fishermen were blamed to infer how strong fisherman C must have been. In cases where the models qualitatively disagree, participants' choices are better accounted for by the rationality model (see trials 5 to 7). Generally these situations were ones like trial 6 discussed above, where the unknown factor implied either that the fisherman acted unreasonably but it didn't end up mattering, or that they acted reasonably but happened to end up being pivotal to the group's success. There were also a number of situations like trial 8 for which participants' selections deviated from our models' predictions. We will return to these in the General Discussion.

Table 1 shows the best-fitting parameters of the different models, together with a measure of fit. Maximum likelihood estimation was performed by grid search, with the two softmax parameters (β_r and β_d) ranging from 0.25 to 9, the k-level reasoning parameter from 1 to 3, and the weight w in the mixture model that takes into account both rationality and pivotality from 0 to 1. The mixture model used the best fitting

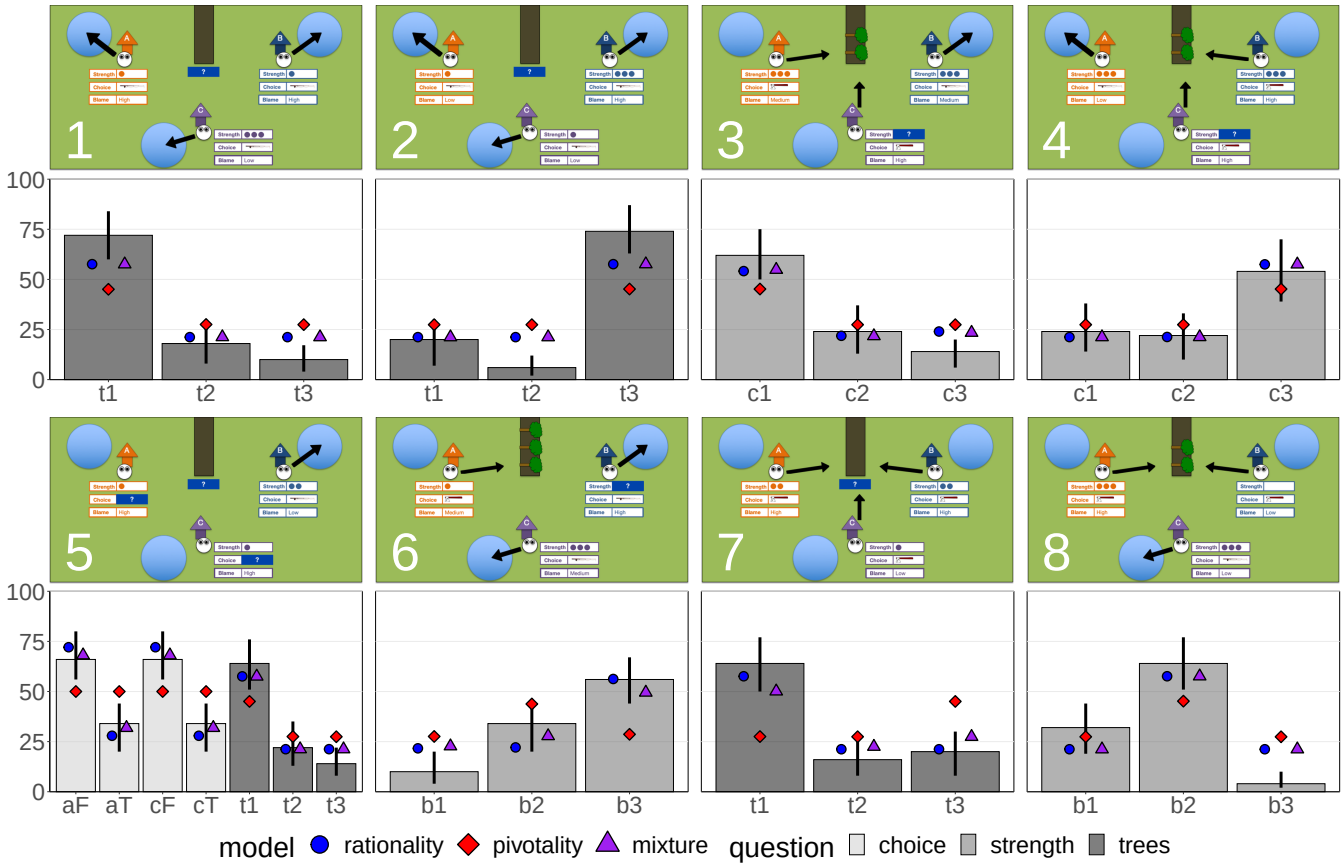


Figure 2: Participants' selections for a subset of trials. Bars show percentage selected with 95% bootstrapped confidence intervals. The symbols show the predictions of the rationality model (blue point), pivotality model (red diamond), and mixture model (purple triangle). *Note*: a, b, c stand for the three fishermen; T = going for the trees, F = going fishing, t = trees. For example, in trial 1, participants had to infer how many trees there were. In trial 3, the missing piece was fisherman c's strength. In trial 5, participants had to infer fishermen a and c's actions, and how many trees there were.

k and β_r parameters for the rationality model, and fit w and β_d. As a baseline, we included a random model that predicts that each response option is chosen with equal probability.

Figure 3 shows the correlation between model predictions and the probability with which participants' selected each response option across all 36 trials (45 total judgments) in the experiment. The rationality model and the mixture model correlate better with participants' responses than the pivotality model. Even after penalizing for the additional free parameters in these models, they provide a better account of participants' responses (see Table 1). While the mixture model has a higher correlation than the rationality model and lower Bayesian Information Criterion value than the rationality model, the marginal difference and high correlation between the models' predictions ($r(\text{mixture}, \text{rationality}) = .96$) suggests that the rationality model accounts for the majority of variance in participants' responses. As Table 1 shows, the best-fitting version of the mixture model places most of its weight on the rationality component ($w = 0.9$).

Table 1: Model comparison. Columns w, k, β_r, and β_d show the best-fitting parameters for each model determined via maximum likelihood estimation. For the Bayesian Information Criterion (BIC) lower values indicate better fit.

model	w	k	β_r	β_d	BIC
Random baseline					3,931
Pivotality				1.5	3,841
Rationality		2	1	1.5	3,653
Mixture	.9	2	1	1.5	3,652

w = the mixture model's weight on the rationality component (Eq. 4)
k = depth of recursion in the k-level reasoning model (Eq. 1)
β_r = softmax parameter in the fishermens' decision function (Eq. 1)
β_d = softmax parameter in the participants' decision function (Eq. 6)

General Discussion

Humans are sophisticated detectives: from scant physical evidence, they can recreate the causal history of events (Gerstenberg, Siegel, & Tenenbaum, 2018; Smith & Vul, 2014; Lopez-Brau et al., 2020; Pelz et al., 2020; Schachner & Kim, 2018; Chen & Scholl, 2016). Here, we show how social evaluations

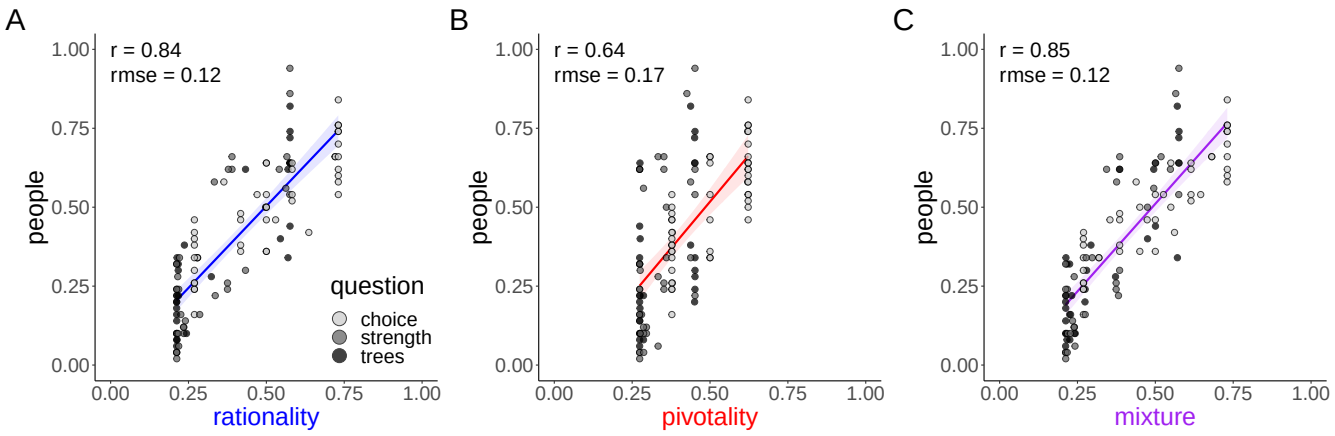


Figure 3: Scatter plots showing how well the the rationality model (A), pivotality model (B), and mixture model (C) account for people's inferences. The shading of the points indicates what kind of information was missing: a fisherman's choice (light), their strength (medium), or the number of trees (dark). r = Pearson's correlation coefficient, rmse = root mean squared error.

provide a rich source for making inferences about what happened. Judgments of blame reveal what actions agents took, what capacities they had, and what the situation was like.

We studied inferences from blame in the fisherman paradigm in which a group of agents failed to coordinate their actions to achieve an optimal outcome. The more blame a fisherman received in our setting, the more likely participants were to infer that he must have failed to do the right thing. This provides converging evidence for the model introduced by Allen et al. (2015) in which attributions of blame are sensitive to normative expectations about how a person should act. Allen et al. (2015) also found that participants' blame judgments were affected by the causal role that a fisherman's action played. Fishermen were blamed more when their action was pivotal, that is, when the optimal outcome could have been achieved had they acted differently.

We found that the inferences that people made from blame judgments mostly reflected whether the agent did the right thing (the rationality component of the model), and were less affected by the causal role that their action played (the pivotality component). This may be the result of a number of different features of the inference process. For one, computing pivotality is computationally demanding. In Allen et al.'s (2015) setting, participants saw what actually happened so computing pivotality was relatively straightforward. However, in our setting, participants have to consider several possibilities and then compute what the agent's causal role would have been in each situation. Computing the rationality of an action, in contrast, is less demanding. For example, evaluating whether an agent did the right thing doesn't require considering what actions the other agents took. The fact that there is a slight mismatch between how people assign blame (pivotality matters), and the inferences that people make based on these judgments (pivotality matters less), suggests that people may sometimes draw the wrong inferences about what happened from others' social evaluations.

Overall, our model did a good job of capturing participants' inferences (see Figure 3). However, for some trials the model's predictions were off. For example, in trial 8 in Figure 2, the rationality model captures participants' belief that the most likely situation was one in which fisherman B's strength was 2. However, it doesn't capture participants' preference $S(B) = 1$ over $S(B) = 3$. The rationality model's ambivalence between these two situations stems from a risk/reward trade-off in the two situations. If $S(B) = 1$, then it never helps the group if he goes for the trees but it also doesn't hurt much (only one fish sack would be wasted). If $S(B) = 3$, then it's possible that going for the trees can help the group (assuming both fishermen A and C decided to go fishing), but this situation is unlikely to arise (so three fish sacks could be wasted). The fact that participants were more likely to infer that fisherman B's strength was 1 rather than 3, suggests that they would have expected that fisherman to receive high blame if three fish instead of one had been wasted.

This paper takes a first step toward looking into how social evaluations support inferences about what happened. We have only considered a small subset of the factors that are known to influence attributions of blame (Malle, 2021; Alicke et al., 2015). In future work, we will look into how people make inferences about an agent's mental states from social evaluations. For example, when two agents took the same action but only one got blamed, we can infer that their mental states were different: one intended to take the action whereas the other acted accidentally (Young & Saxe, 2011). Combining physical evidence and social evaluations licenses inferences about an agent's capacities. When two agents achieved the same outcome but one is praised more than the other, this suggests that one agent exceeded the expectations whereas the other just met them. Social evaluations provide a rich source of information about what happened. A source that we regularly draw on in our everyday lives, and one that we only begin to understand scientifically.

Acknowledgments

We thank the members of the Causality in Cognition Lab (CiCL) for feedback and discussion.

References

Alicke, M. D., Mandel, D. R., Hilton, D., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives on Psychological Science*, *10*(6), 790–812.

Allen, K., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M., & Tenenbaum, J. B. (2015). Go fishing! responsibility judgments when cooperation breaks down. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 84–89).

Chen, Y.-C., & Scholl, B. J. (2016). The perception of history: Seeing causal history in static shapes induces illusory motion perception. *Psychological Science*, *27*(6), 923–930.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*(1), 93–115.

Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. B. (2018). What happened? reconstructing the past from vision and sound. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (p. 409).

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122-141.

Kirfel, L., Icard, T. F., & Gerstenberg, T. (2020). Inference from explanation. *PsyArXiv*.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.

Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2020). Mental state inference from indirect evidence through bayesian event reconstruction. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 467–473).

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318.

Pelz, M., Schulz, L., & Jara-Ettinger, J. (2020). The signature of all things: Children infer knowledge states from static images. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (p. 1977).

Schachner, A., & Kim, M. (2018). Alternative causal explanations for order break the link between order and agents. *PsyArXiv*. (https://psyarxiv.com/jd2qr/)

Smith, K., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 1467–1472).

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548.

Wu, Y., Schulz, L., Frank, M. C., & Gweon, H. (2021). Emotion as information in early social learning. *PsyArXiv*.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12), e1000254.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202-214.