Inference from social evaluation

Zachary J. Davis[1], Kelsey R. Allen[2], Max Kleiman-Weiner[3], Julian Jara-Ettinger[4], and Tobias Gerstenberg[1]

[1]Stanford University

[2]Google Deepmind

[3]Washington University

[4]Yale University

Abstract

People have a remarkable ability to infer the hidden causes of things. From physical evidence, such as muddy foot prints on the floor, we can figure out what happened and who did it. Here, we investigate another source of evidence: social evaluations. Social evaluations, such as praise or blame, are commonplace in everyday conversations. While such evaluations don't fully reveal what happened, they provide valuable clues. Across three experiments, we present situations where a person was praised or blamed, and participants' task is to use that information to figure out what happened. In Experiment 1, we find that people draw systematic inferences from social evaluations about situational factors, a person's actions, capabilities, and social roles. In Experiments 2 and 3 we develop computational models that generate praise and blame judgments by considering what causal role a person's action played, and what action they should have taken. Inverting these generative models of praise and blame via Bayesian inference yields accurate predictions about what inferences participants draw based on social evaluations.

Keywords: social evaluations; causality; inference; social cognition; blame.

To appear in Journal of Personality and Social Psychology: Attitudes and Social Cognition

*Corresponding author: Tobias Gerstenberg https://orcid.org/0000-0002-9162-0779, Stanford University, Department of Psychology, 450 Jane Stanford Way, Bldg 420, Stanford, CA 94305, Email: gerstenberg@stanford.edu. All the data, study materials, pre-registrations, and analysis code are available here: `https://github.com/cicl-stanford/inference_from_social_evaluation`

Statement of limitations

We looked at attributions of blame and praise, and what inferences people can draw from such attributions. As there is no clear normative standard for how such judgments and inferences should be made, we weren't able to bonus participants based on performance. We assume that participants were sufficiently engaged in the task to produce systematic judgments. Participants acted as third-party judges in hypothetical scenarios, rather than producing behaviors that would be directly consequential to themselves or others. Across three experiments, we only considered a limited set of situations, and our experimental paradigms abstracted away many factors that make responsibility judgments challenging in the real world. We developed computational models of praise (in Experiment 2) and blame (in Experiment 3). While these models capture much of the variance in participants' judgments, it's possible that better models exist. Our models assume that people are making sophisticated inferences. However, it's possible that participants arrived at their judgments using heuristic shortcuts. All of our experiments relied on adult participant samples recruited online through MTurk and Prolific. This limits the potential generalizability of our findings.

Introduction

Humans are evaluative creatures. Social evaluations, such as attributions of blame and praise, form an important part of our daily lives (Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Anderson, Crockett, & Pizarro, 2020; Lagnado, Gerstenberg, & Zultan, 2013; Malle, 2021). Many factors influence how we evaluate others. For example, we blame a person more for a negative outcome when they acted intentionally and knowingly (Kirfel & Phillips, 2023; Lagnado & Channon, 2008; Malle, Guglielmo, & Monroe, 2014). When the person we blame is a friend, we expect them to make up for it. Indeed, relationship regulation may be one of the key functions of moral cognition (Rai & Fiske, 2011; Sarin, Ho, Martin, & Cushman, 2021). We generally aim to avoid people who have slighted us in the past, and to strengthen our ties with those who have supported us (e.g., Barclay & Willer, 2007; Baumard, André, & Sperber, 2013).

Sometimes we directly blame or praise others and sometimes we share our social evaluations with others. Such social evaluations provide a rich source of information about what happened. When we communicate with one another, we leave many things unsaid. Nonetheless, we generally have no trouble understanding what happened because we know how to fill in the gaps (Kirfel, Icard, & Gerstenberg, 2022). For instance, imagine that Alice told Bob that the goalie was to blame for her favorite soccer team's loss. Based on what Alice said, Bob, who hasn't seen the game, might infer that the goalie failed to save an easy shot. If Alice had blamed the same team's striker instead, Bob would have inferred that the striker missed an easy goal.

People are very good at drawing inferences that go beyond what can be perceived directly (Beller, Xu, Linderman, & Gerstenberg, 2022; Gerstenberg, Siegel, & Tenenbaum, 2021; Smith & Vul, 2014; Wu et al., 2024). Based on physical evidence, people can infer what actions someone took ("my roommate left the fridge open"; e.g., Lopez-Brau, Kwon, & Jara-Ettinger, 2022; Schachner & Kim, 2018), what their goals were ("they must have been hungry"; e.g., Baker, Tenenbaum, & Saxe, 2006), and what they knew ("they thought that we still had some leftover food"; e.g., Pelz, Schulz, & Jara-Ettinger, 2020). Clues about what happened reside not only in the physical world. They also inhabit the people around us. A person's emotional expression, for example, can reveal what happened (Ong, Zaki, & Goodman, 2019; Saxe & Houlihan, 2017; Weiner, 1985; Wu, Schulz, Frank, & Gweon, 2021). Even infants can make these inferences. Upon hearing a "Whoa?" or an "Aww!" from an adult, they infer that the person must have looked at a fancy toy or a cute baby (Wu, Muentener, & Schulz, 2017). And when two people disagree with one another, children infer that they are talking about something ambiguous (Amemiya, Heyman, & Gerstenberg, 2024; Amemiya, Walker, & Heyman, 2021).

These various inferences have something in common: they rely on people's intuitive understanding of how the physical world works and how people work. This intuitive understanding allows us to reason not only from cause to effect, but also from effect to cause (Gerstenberg & Tenenbaum, 2017). For example, if someone really likes coffee (the cause), they will walk far to get one if they need to (the effect). In turn, having seen someone walk a long distance to get a coffee (the effect) tells us that they must really like it (the cause; e.g., Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). In this paper, we study how people make inferences about what happened from social evaluations. Because systematic

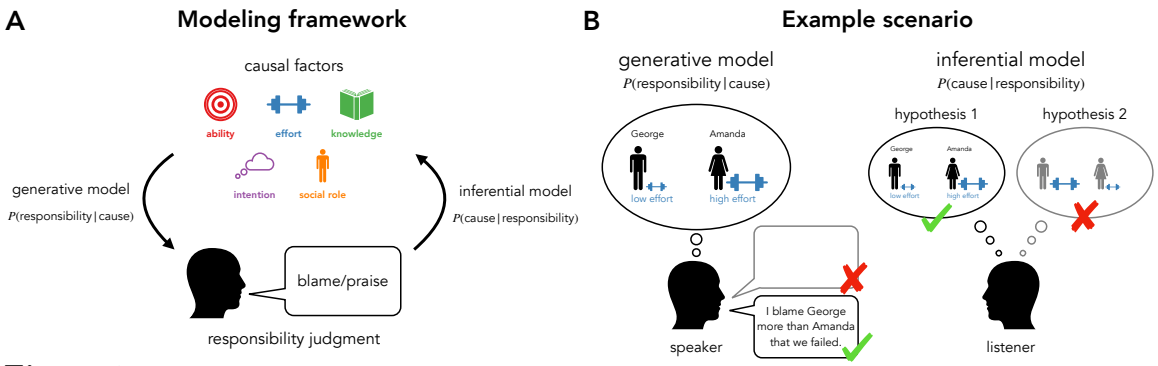


Figure 1

Illustration of the overall modeling framework (A) and a specific example scenario (B). ***A*** *Several causal factors systematically influence how we hold others responsible for the outcomes of their actions. The* generative model *captures how much responsibility a person receives depending on these causal factors. The* inferential model *captures what causal factors were likely to be true given how much responsibility a person received.* ***B*** *The speaker uses a generative model to reason from the causal factors to the responsibility judgment $P(\text{responsibility} \mid \text{cause})$. Here, they choose which of two responsibility judgments to produce based on what happened. Under the model, since George put in less effort than Amanda, the speaker blames George more than Amanda. The listener uses Bayesian inference to invert the listener's generative model and reasons from the responsibility judgment to the causal factors $P(\text{cause} \mid \text{responsibility})$. In this example, the listener infers based on the speaker's utterance, that George was more likely to have put in less effort than Amanda (which is what caused the speaker to blame George more).*

factors underlie people's social evaluations, and people share intuitions about what makes a person blameworthy (or praiseworthy), knowing who got the blame (or praise) provides information about what happened (see Figure 1a). If a person received a lot of praise for picking up a coffee, you can infer that they must have put in a lot of effort to get it.

The paper is structured as follows. First, we motivate our studies by discussing existing theoretical frameworks for social evaluations and related work on how people may draw inferences from such evaluations. Then, we present a formal modeling framework that makes Bayesian inferences from responsibility judgments about the different factors that gave rise to these judgments. We evaluate this modeling framework in three experiments that look at inferences in increasingly complex scenarios. Experiment 1 employs short vignettes in which one person received more (or less) blame than another, and participants' task is to infer the persons' attributes (which person put in more effort, was more able, knew more, etc.). Experiment 2 goes beyond simple binary judgments and asks participants to infer what path a person must have traveled to the office based on the praise they received from their coworker for picking up a coffee. Here, we compare participants' graded inferences to the predictions of a computational model that links the praise a person receives to the amount of effort they incurred. Finally, Experiment 3 looks at a multi-agent setting where participants are asked to infer what happened based on the degree of blame that each agent

received. Again, we compare participants' inferences against the predictions of different computational models that link social evaluations to whether a person did the right thing, and if not, how much that mattered. Overall, we find that participants are capable of drawing sophisticated inferences about what happened from responsibility judgments. They can do so because people share systematic intuitions about what makes someone responsible. We conclude by highlighting the main insights from this work, discussing limitations and future directions, as well as broader implications.

Models of responsibility judgments

Prior research has uncovered systematic factors that influence people's responsibility judgments. These factors can be broadly classified into two main components: factors about the causal role that the agent played in bringing about the outcome, and factors about what an action reveals about the kind of person they are (Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015; Davis, Allen, & Gerstenberg, 2021; Gerstenberg, 2024; Gerstenberg et al., 2018; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021). For a person to be held responsible, their action must have made a difference to the outcome (e.g., Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Zultan, Gerstenberg, & Lagnado, 2012). And we hold others more responsible for negative outcomes when their actions reflect a bad moral character (e.g., Carlson, Bigman, Gray, Ferguson, & Crockett, 2022; Hamilton, 1978; Schlenker, Britt, Pennington, Murphy, & Doherty, 1994; Uhlmann, Pizarro, & Diermeier, 2015).

That said, we are still far from a comprehensive understanding of how people attribute responsibility to one another. Early theories of responsibility judgments took the form of decision-stage models. For example, Shaver (1985) proposed a normative theory of blame attribution according to which one first needs to establish a person's causal role, then assign responsibility based on factors that include intentionality and foreseeability, and finally attribute blame if there were no justifications or excuses. Other decision-stage models like Weiner's (1995) highlight the role that controllability plays for responsibility attributions (see also Alicke, 2000). Malle et al. (2014) propose a path model of blame in which intentionality plays a critical role. If an action was intentional, a person's reasons for acting are considered. If the action wasn't intentional, their obligation and capacity to have prevented the negative outcome from happening matter (Sarin & Cushman, 2024). These frameworks all postulate a similar process for assigning responsibility (or blame). A causal analysis comes first, and then additional factors are considered that determine whether the person is an appropriate target for responsibility.[1]

A limitation of these models is that they don't make fine-grained predictions about how responsible a person will be held. Computational models of responsibility provide such predictions, focusing on the causal role that an action played for the outcome. For example, some models explain responsibility judgments as reflecting how much a person's actions changed the subjective probability of the outcome (Brewer, 1977; Fincham & Jaspars, 1983;

[1] Alicke's (2000) culpable control model predicts a different sequence of events. It suggests that the desire to blame someone kicks in early and biases the causal analysis of what happened (see also Knobe, 2010; Sytsma, 2021).

Johnson & Rips, 2015; Spellman, 1997). Others use counterfactuals to capture the extent to which a person's action made a difference (Chockler & Halpern, 2004; Engl, 2022; Felsenthal & Machover, 2009; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Naumov & Tao, 2018; Quillien & Lucas, 2023). A person is more responsible the closer their action was to making a difference to the outcome (Chockler & Halpern, 2004; Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Zultan et al., 2012).

More recent computational models of responsibility have incorporated inferences about the person in addition to their causal role (Gerstenberg, 2024; Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa et al., 2021; Wu, Sridhar, & Gerstenberg, 2023). Gerstenberg et al. (2018) showed that judgments of blame and praise were reflective of how pivotal an action was, as well as what that action revealed about the person. For example, when a goalkeeper saved an unexpected shot they received praise, whereas when a gambler predicted an unexpected outcome they didn't. The goalie's action was indicative of skill whereas the gambler made a bad decision and just got lucky. Langenhoff et al. (2021) showed that causal attributions and dispositional inferences affect responsibility judgments for voting outcomes. Responsibility judgments were influenced both by how close a committee member's vote was to making a difference to the outcome, as well as whether their vote was unexpected given their party affiliation. Similarly, Kleiman-Weiner et al. (2015) showed that people's judgments in moral dilemmas were sensitive both to inferences about the person's intentions as well as what difference their action made (see also Halpern & Kleiman-Weiner, 2018). Sosa et al. (2021) demonstrated that participants' moral responsibility judgments about agents in animated scenarios are also sensitive to what causal role their action played and what the action revealed about their character. In their study, Sosa et al. captured the causal role computationally by simulating what would have happened in a relevant counterfactual situation in which the agent hadn't been there (see Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021). For the person inference part, the model considered how much effort the agent exerted which is indicative of how much they desired the harmful outcome (see Jara-Ettinger et al., 2016). And, finally, Wu et al. (2023) showed that when judging responsibility in dynamic helping and hindering scenarios, participants cared about what causal role the agent played, and what the agent's actions revealed about their intentions (see also Ullman et al., 2009).

Inferences from responsibility judgments

Because of the systematic ways in which people hold others accountable for their actions, responsibility judgments provide a rich source of information. If people's intuitions about how blame should be attributed are generally shared, then knowing how much a person was blamed may reveal what they did. Prior work has looked at people's inferences about what happened based on physical evidence (Beller et al., 2022; Gerstenberg, Siegel, & Tenenbaum, 2021; Lopez-Brau et al., 2022; Smith & Vul, 2014), or based on the emotional expressions of others (Wu et al., 2021). No work to date has looked at what people can learn from responsibility judgments. However, there has been some work on inferences from related concepts: punishment and explanations.

Radkani, Tenenbaum, and Saxe (2022) construe punishment decisions as rational, communicative, social acts (see also Dunlea & Heiphetz, 2020; Sarin et al., 2021). Accordingly, when a person decides to punish, they consider the cost to the target of the pun-

ishment, the potential social benefit (e.g., via deterrence), the cost to themselves, and also what reputational costs their action might bear (through the inferences that others would make about them based on their action; see also Ho, Saxe, & Cushman, 2022; Kleiman-Weiner, Shaw, & Tenenbaum, 2017; Yoon, Tessler, Goodman, & Frank, 2020). Radkani and Saxe (2023) use this framework to show that people can draw inferences from punishment about the wrongness of an action, and about the punisher's motivations. For example, from knowing that a person has a just motivation and chose to punish, an observer can infer that the target's action is likely to have been morally wrong. Conversely, if we know that an action was not wrongful but the target was punished anyhow, we might infer that the punisher is less concerned with justice and more with harming the target.

Kirfel et al. (2022) explored what inferences people draw from others' explanations (see also Navarre, Konuk, Bramley, & Mascarenhas, 2024). Just like there are systematic factors that influence people's responsibility judgments and punishments, the way in which people select explanations is systematic, too. For example, when several events contributed to some outcome, people generally have a preference to select an abnormal event rather than a normal event as the explanation for what happened (Hart & Honoré, 1959/1985; Hilton & Slugoski, 1986; Kahneman & Miller, 1986). The forest caught fire because of the arsonist rather than the presence of oxygen. Recently, it was found that both event normality and causal structure affect people's explanation selections (Gerstenberg & Icard, 2020; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). In conjunctive causal structures, where each cause is necessary to bring about the outcome, people select abnormal events as causes. However, in disjunctive structures, where each cause is individually sufficient to bring about the outcome, people select *normal* events as causes. Given that people have systematic preferences for selecting explanations, what explanation someone gave provides information about what happened. For example, imagine that you knew your friend had to pass biology and chemistry to get into medical school. If they told you "I got into medical school because I passed chemistry!", you can infer that, for them, chemistry was the more difficult exam to pass (Kirfel et al., 2022). Similarly, when a listener knows about the normality of each event, they can infer the underlying causal structure from someone's explanation. For example, if someone cites the normal event as the cause, one can infer that the structure must have been disjunctive.

While there is some overlap between the factors that drive punishment, explanation selection, and responsibility judgments, these concepts are nonetheless distinct. Here, we look at what inferences people draw from knowing how much a person was blamed or praised.

Modeling inferences from responsibility judgments

Consider the following scenario:

> "Andrew was assigned to do a group project with Amanda and George. Their group did a bad job and received a failing grade for the assignment. Andrew blamed George more than Amanda for the group's failing grade. Who didn't try very hard on the assignment?"

Was it Amanda or George? The answer is intuitive: George didn't try very hard. Figure 1b illustrates the reasoning process that underlies this intuitive inference. We begin

with a *generative model* of how people assign blame (or praise) depending on the context. In this example, we assumed a simple generative model of blame: one where low effort generates high blame, and high effort generates low blame. The listener uses an *inferential model* to invert the speaker's generative model. The listener considers two possibilities: George didn't try hard but Amanda did (hypothesis 1), or George tried hard but Amanda didn't (hypothesis 2). The listener uses the inferential model to update their belief about each hypothesis based on Andrew's responsibility judgment (i.e., more blame was assigned to George than to Amanda). The listener considers in which of the two possible situations the speaker would have been more likely to say what they did, and update their beliefs proportional to the likelihood. Based on the speaker's responsibility judgment, it's more likely that George didn't try as hard as Amanda (hypothesis 1).

We propose that these intuitions, where we reconstruct what happened by working backwards through a causal model of blame, can be captured via Bayesian inference. In this example, the listener considers two candidate hypotheses, and infers their probability by conditioning on the information provided by the speaker. Formally,

$$p(\text{cause} \mid \text{responsibility}) = \frac{p(\text{responsibility} \mid \text{cause}) \cdot p(\text{cause})}{\sum_{i=1}^{n} p(\text{responsibility} \mid \text{cause}_i) \cdot p(\text{cause}_i)} \tag{1}$$

Here, $p(\text{cause})$ is the prior over the different possible causes, and $p(\text{responsibility} \mid \text{cause})$ is the likelihood that a speaker would produce a certain responsibility judgment if a particular cause was true.[2]

This example highlights two key aspects of our proposal. First, the exact content of the generative model and what inferences can be drawn are context sensitive. Different situations inevitably involve different sets of causes, different hypothesis spaces about what might have happened, and different ways in which people's behavior combines to produce outcomes. Second, this context sensitivity is held together by a unified set of principles, where reconstructing what happened involves inverting a generative model through Bayesian inference, using a context-general set of intuitions about what causes people to blame or praise each other.

With this in mind, our goal is to test people's capacity to reconstruct what happened from social evaluations in a variety of different domains. This way, we can test that the same principles apply to a broader set of social situations inspired by everyday life. To accomplish this, Experiment 1 establishes the general phenomenon, using vignette scenarios like the one with Amanda and George, manipulating several factors that have been shown to influence responsibility judgments. Experiments 2 and 3 then test this capacity in two different, complementary paradigms. This helps highlight both how people adapt the exact generative models to the situation at hand, while also revealing the core set of principles in how we reconstruct causes from social evaluations.

In Experiment 1, the generative model simply maps from a given factor (e.g., low effort or little knowledge) to a responsibility judgment (e.g., high blame or low blame; see

[2] For example, let's assume that the listener considers each hypothesis equally likely a priori $p(\text{cause} = \text{`hypothesis 1'}) = p(\text{cause} = \text{`hypothesis 2'}) = 0.5$, and that a speaker is likely to blame George more when George put in low effort, $p(\text{responsibility} = \text{`blame George'} \mid \text{cause} = \text{`hypothesis 1'}) = 0.8$, but less likely to do so when George put in high effort $p(\text{responsibility} = \text{`blame George'} \mid \text{cause} = \text{`hypothesis 2'}) = 0.3$. With these assumptions, a listener would then be able to infer how likely it was that George put in low effort, given the he was blamed $p(\text{cause} = \text{`hypothesis 1'} \mid \text{responsibility} = \text{`blame George'}) = \frac{0.8 \cdot 0.5}{0.8 \cdot 0.5 + 0.3 \cdot 0.5} = 0.72$.

Figure 1b). In Experiments 2 and 3, we consider a subset of these factors but manipulate them in more fine-grained ways which allows us to make precise quantitative predictions about what responsibility judgments a speaker is likely to produce, and what inferences a listener is likely to draw. Inspired by prior work, these generative models assume that people's responsibility judgments are sensitive to what causal role a person's action played, and what the action revealed about them (e.g. Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa et al., 2021). In Experiment 2, we hold the person's causal role constant but manipulate what their action reveals about them. In Experiment 3, we manipulate both the person's causal role and what the action reveals about them.

For each experiment, we present concrete model implementations which can be viewed as special cases of the more general framework. While the implementation details of these generative models are adapted to the specifics of each experimental setting, the inferential model largely stays the same. As illustrated in Figure 1a, we assume that the listener uses Bayesian inference to figure out what happened by considering what situation would have made the speaker say (or judge) what they did.

Experiment 1: Inference from blame about different factors

The goal of Experiment 1 is to establish that people can use social evaluations to make inferences about what happened. Several factors affect how people attribute responsibility. In this experiment, we looked at five: a person's ability (Gerstenberg, Ejova, & Lagnado, 2011; Guglielmo & Malle, 2010), how much effort they put in (Bigman & Tamir, 2016; Jara-Ettinger et al., 2016; Sosa et al., 2021), what they knew and didn't know (Gerstenberg & Lagnado, 2012; Kirfel & Lagnado, 2019; Kirfel & Phillips, 2023; Lagnado & Channon, 2008), what they intended (Kleiman-Weiner et al., 2015), and what social role they had (Hamilton, 1978; McManus, Kleiman-Weiner, & Young, 2020; Rai & Fiske, 2011; Schlenker et al., 1994). We briefly discuss prior research on how these five factors influence responsibility judgments.

Factors influencing responsibility judgments

Ability

How much praise or blame a person receives depends on their ability. For example, when a more capable person refuses to help, they are blamed more than a less capable person (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015). Similarly, a more capable person receives more blame for failing in an achievement task (Gerstenberg et al., 2011).

Effort

How much effort a person exerted also matters for how responsible we hold them. Putting in little effort results in more blame for failures, and putting in a lot of effort results in more praise for successes (Bigman & Tamir, 2016; Sosa et al., 2021; Weiner & Kukla, 1970; Xiang, Landy, Cushman, Vélez, & Gershman, 2023). Because effort is costly, putting in effort indicates a strong desire for the outcome. When the outcome is negative, such as another agent being harmed, putting in more effort leads to *more* blame. Assuming that others take into account the expected costs and rewards of their actions (Jara-Ettinger et

al., 2016), the fact that a person was willing to put in a lot of effort indicates that they strongly desired to bring about the harmful outcome (Sosa et al., 2021).

Knowledge

When we hold others responsible, it not only matters what they were capable of doing, and what they did, but also what their actions reveal about their mental states. For example, knowledgeable people are generally held more responsible for the outcomes of their actions than ignorant people (Gilbert, Tenney, Holland, & Spellman, 2015; Kirfel & Lagnado, 2021; Kirfel & Phillips, 2023; Lagnado & Channon, 2008). Bringing about an outcome unknowingly, doesn't reveal much about the underlying desires (but see Hertwig & Engel, 2016; Kirfel, Bunk, Zultan, & Gerstenberg, 2023; Young & Saxe, 2011). In contrast, when a person knows the consequences of their action, this licenses an inference that they must have wanted for that outcome to come about, or at least that they were okay with it happening (e.g., when outcome was a foreseeable but unintended side-effect of their action; Kleiman-Weiner et al., 2015; Knobe, 2010; Sloman, Fernbach, & Ewing, 2012).

Intention

In addition to what someone knew, it also matters what they intended. We blame others more for harms they brought about intentionally versus accidentally (Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009; Gerstenberg, Lagnado, & Kareev, 2010; Lagnado & Channon, 2008; Malle et al., 2014).

Social role

Finally, when attributing responsibility, it matters what we expected from others. A person's role (Hamilton, 1978; Schlenker et al., 1994) and their relationship to us (Kleiman-Weiner, Saxe, & Tenenbaum, 2017; McManus et al., 2020; Powell, 2022; Rai & Fiske, 2011) affects our expectations. For example, from a close friend, we expect that they care about us and that they're willing to support us when we need them. So, we would blame a closer friend more for not helping us when in need compared to a more distant friend (Scholten, 2022).

Methods

For each of the five factors (ability, effort, knowledge, intention, and social role), we created three scenarios using a structure similar to the one with Andrew, Amanda, and George mentioned above. In each scenario, person A blames person B more (or less) than person C, and participants' task is to figure out what happened. We pre-registered the sample size, experiment materials, and statistical analyses here: https://osf.io/yrjd6.

Participants

50 participants (*age*: M = 26, SD = 7; *gender*: 24 female, 25 male, 1 non-binary; *race*: 1 African American/Black, 6 Asian American/Asian, 36 White/Caucasian, 7 other) were recruited via Prolific. Participants were at least 18 years old and had been approved on more than 90% of previous tasks. They were paid $1.50.

Design & Procedure

All participants read 16 vignettes including one attention check. The vignettes were presented in randomized order, but the attention check always appeared as the 8th trial.[3]

Table 1 shows five example vignettes. Each vignette described a background scenario in two to four sentences (e.g., "Andrew was assigned to do a group project with Amanda and George. Their group did a bad job and received a failing grade for the assignment."). Then one of the characters is blamed more or less (e.g., "Andrew blamed George more than Amanda for the group's failing grade."). Finally, a question was posed (e.g., "Who tried very hard on the assignment?"). In each scenario, we randomized whether the character was blamed *more* or *less*, and we randomized the directionality of the question (e.g., "tried very hard" vs. "didn't try very hard"). The names of the characters in the vignettes were randomly sampled without replacement from the top 25 male and female names over the last 100 years in the United States. You can find a full list of vignettes in Appendix A. Participants responded by clicking one of two buttons which showed the name of each character (e.g., "George" versus "Amanda"). It took participants on average 7.1 minutes (SD = 2.9) to complete the task.

Transparency and Openness

All data have been made publicly available and can be accessed at `https://github.com/cicl-stanford/inference_from_social_evaluation`.

Hypotheses

Based on prior research having shown what factors systematically influence attributions of blame, we have the following hypotheses about what inferences participants draw based on information about who was blamed more (or less). Assuming that Person 1 was blamed *more* than Person 2, we predict that participants are more likely to select Person 1 as the one who was more able to help (ability), put in less effort (effort), knew that something bad was going to happen (knowledge), intended for the bad outcome to happen (intention), and was more closely related to the affected person (social role).

Alternatively, it could be the case that even though these factors affect how people assign blame, people won't make systematic inferences about them from knowing how much a person was blamed. In that case, participants should be no more likely to select one person or the other.

Results

Figure 2 shows the proportion of participants who selected the person in line with our predictions, separately for each scenario. To test our main hypothesis, we ran a Bayesian logistic mixed effects model with a fixed intercept, random intercepts per participant, and nested effects of scenario within scenario type.[4] We coded the responses such that 1 responded to the predicted person in the scenario, and 0 to the other person. We also re-coded

[3] The attention check describes which of two characters mowed the lawn and asked participants "Who mowed the lawn?". No participant failed the attention check.

[4] All Bayesian models reported in this paper were written in Stan (Carpenter et al., 2017) and accessed with the `brms` package (Bürkner, 2017) in `R` (R Core Team, 2019).

Table 1

Example scenario for each of the five different factors that participants had to infer. Each scenario features three people: A Requester who assigns more or less blame to Person 1 compared to Person 2. Note*: For each Factor, the experiment included three scenarios, with 15 scenarios in total. Figure 2 shows the results for the indexed scenarios.*

Factor	Index	Scenario
Ability	1	*Requester* needed help fixing their flat tire. *Requester* asked *Person 1* and *Person 2* for help, but neither did. *Requester* didn't manage to fix the tire by themselves. *Requester* blamed *Person 1* more/less than *Person 2* for not helping. Who is better/worse at fixing tires?
Effort	4	*Requester* was assigned to do a group project with *Person 1* and *Person 2*. Their group did a bad job and received a failing grade for the assignment. *Requester* blamed *Person 1* more/less than *Person 2* for the group's failing grade. Who tried / didn't try very hard on the assignment?
Knowledge	7	*Requester* texted *Person 1* and *Person 2* to pick them up from the airport. Neither responded to their text. *Requester* later learned that only one of them saw their message. *Requester* blamed *Person 1* more/less than *Person 2* for not helping. Who saw / didn't see the text message?
Intention	10	*Requester* fell over while playing soccer with *Person 1* and *Person 2*. Both *Person 1* and *Person 2* stepped on *Requester*. *Requester* blamed *Person 1* more/less than *Person 2* for stepping on them. Who stepped on *Requester* on purpose / by accident?
Social role	13	*Requester* had a birthday party. *Requester* asked *Person 1* and *Person 2* to come, but neither came. *Requester* blamed *Person 1* more than *Person 2* for not coming. Who is *Requester*'s closer / less close friend?

the data based on whether we used the "blamed more" or "blamed less" frame, and based on how the question was posed (e.g., who was the closer vs. less close friend). As predicted in our pre-registration, the fixed intercept in the model is positive and the 95% credible interval excludes 0, $\beta = 2.40$ (95% credible interval $[1.74, 3.15]$). This means that overall, participants were more likely to select the predicted person across the different scenarios.

In addition to our pre-registered analysis, we also found that participants not only had an overall preference for selecting the predicted person across all of the scenarios, but that this preference showed up in each of the 15 scenarios. This effect was consistent across participants. On average, participants selected the predicted person 13.16 times out of 15 times (95% confidence interval [12.62, 13.66]).

Discussion

Experiment 1 shows that participants are capable of drawing inferences about unobserved information based on social evaluations. We looked at situations in which someone blamed one person more than the other. Based on this information, participants inferred who was more capable, who put in more effort, who knew more, who intended for the out-

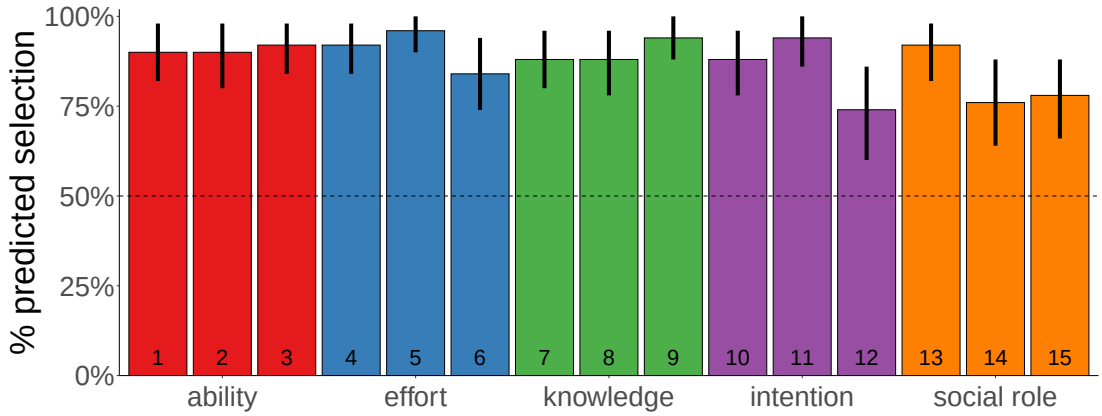


Figure 2

Experiment 1 – Probability with which participants selected the predicted character in each of the fifteen scenarios. *Colors indicate the scenario type. See Table 1 for scenario examples (the indices at the bottom of the bars match the 'Index' column in the table), and Appendix A for the full list of scenarios.* Note*: Error bars show bootstrapped 95% confidence intervals.*

come to happen, and who was more closely related. While prior work has established that manipulating these factors systematically affects people's blame judgments, here we show that people can draw inferences about these factors based on who was blamed. Overall, participants' selections were consistent with our hypotheses. The majority of participants selected the predicted character in each of the fifteen scenarios.

Experiment 2: Inference from praise for doing a favor

Experiment 1 showed that people can draw inferences about several factors based on blame judgments. However, the results of Experiment 1 were only qualitative – we were able to predict which of two characters participants would select, but we weren't able to make quantitative predictions about the strength of their preference. In Experiment 2, we develop and test a computational model that makes quantitative predictions about people's social evaluations, and the inferences they draw based on others' evaluations. Because the model makes quantitative predictions, we can test the key idea more directly that people infer what happened from others' social evaluations by considering what social evaluations they themselves would give in different situations. In this experiment, we focus on the role that effort plays in praise judgments and inferences.

As mentioned before, prior work has shown that how much effort someone exerts affects the extent to which they are held responsible for an outcome (Bigman & Tamir, 2016; Sosa et al., 2021; Weiner & Kukla, 1970; Xiang et al., 2023). Exerting effort is costly, so a person who was willing to incur a cost to bring about an outcome must have desired for that outcome to happen. Assuming that people plan and choose actions so as to maximize their expected utility, the more effort a person exerts, the stronger their desire for the outcome must have been (see Jara-Ettinger et al., 2016). This means that if the outcome is positive, they deserve praise for demonstrating that they value the outcome. If the outcome

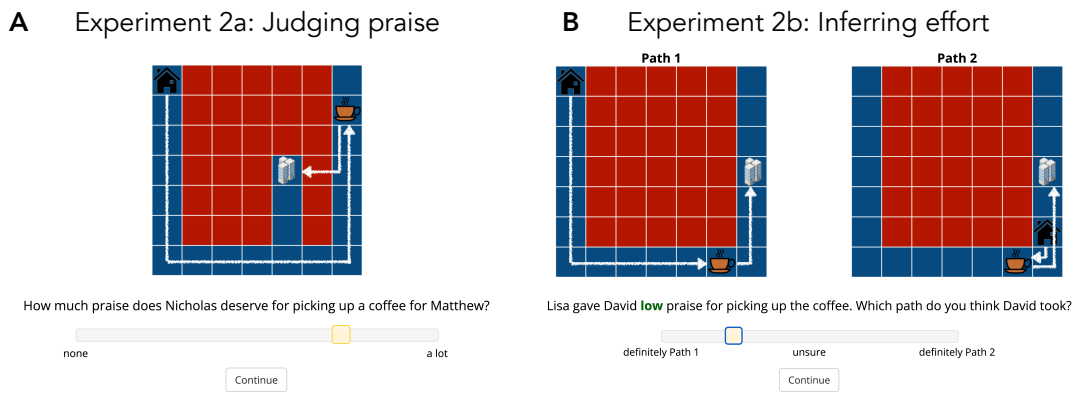


Figure 3

Experiment 2 – Screenshots of the main task. ***A*** *In Experiment 2a, participants judge how much praise one character deserves for picking up a coffee for a co-worker. The black house indicates where the character lives, the coffee indicates the coffee shop, and the white building indicates the office. The white arrows show what path the character took to pick up the coffee on the way to the office. Red tiles indicate busy areas that the cost of moving on these tiles is higher.* ***B*** *In Experiment 2b, participants had to infer from someone's praise judgment (that could be low, medium, or high) what path the person took.*

is negative, such as causing someone else harm, they deserve to be blamed for demonstrating that they desired the bad thing to happen.

In Experiment 1, we focused on blame as a social evaluation. This time, we look at praise. We develop a computational model that predicts praise judgments based on how much a person desired to do another person a favor. The model infers the strength of that desire from the amount of effort a person was willing to exert. We test the model on stimuli that quantitatively vary how much effort one person exerted to do another person a favor. By inverting the generative model that goes from perceived effort to praise, we can make predictions about how much effort someone exerted from knowing how much they were praised. In Experiment 2a, we test the forward direction of the model, going from effort to praise. In Experiment 2b, we test the inverse direction, going from praise to effort.

Overview of the experimental paradigm

Figure 3a shows a screenshot of the *praise judgment* task in Experiment 2a. Participants view the path that one of the characters, Nicholas in this case, took on his way to the office. In each of the scenarios, the character first goes to a coffee shop to pick up a coffee for their co-worker, Matthew in this case. Participants are asked to evaluate how much praise Nicholas deserves for picking up a coffee for Matthew. The scenarios manipulate the location of the character's home (the black house), the location of the coffee shop (the orange coffee symbol), the location of the office (the white building), as well as which areas are busy (red) and which ones aren't (blue). Busy areas are more costly to pass through.

Figure 3b shows a screenshot of the *effort inference* task in Experiment 2b. This time, participants learn how much praise a person received and their task is to infer which

of two images shows the path that the person took. Here, Lisa gave David "low" praise for picking up the coffee. While Path 1 was longer overall than Path 2, on Path 1 the coffee shop was along the way to the office, while on Path 2, the person had to go out of their way to get the coffee. Between trials, we manipulated how much praise a person received (low, medium, or high), as well as what the two candidate paths looked like.

A computational model of praise based on effort

How can we model the extent to which a person deserves praise for doing a favor? The main idea is that people can infer about how much a person valued something from the costs they were willing to incur to bring it about. Here, the outcome is motivated by a pro-social goal: doing the co-worker a favor by bringing them a coffee. The more costly it was for the person to do that, the more they must have valued their co-worker. In our setting, the cost that a person incurs is equal to the effort it took them to bring about the desired goal.

According to the Naïve Utility Calculus (Jara-Ettinger et al., 2016), people generally assume that others choose goals and actions in order to maximize expected utility. Maximizing utility means to obtain as much reward as possible while minimizing costs. By assuming that others act in such a way, we can draw inferences about their desires (and beliefs) from their actions (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Dennett, 1987; Liu, Ullman, Tenenbaum, & Spelke, 2017). For example, if someone is willing to walk further to get an apple when they could have gotten an orange more easily, we can infer that they must have wanted the apple more than the orange. In contrast, when the apple was easier to get than the orange, we can't be sure that they liked the apple more – it's possible that they liked the orange more but weren't willing to incur the extra cost of having to walk further. The same idea can be applied not only to personal goals (eating an apple) but also to social goals (making someone else happy). For example, one can construe the social goals of helping (or hindering) as placing a positive (or negative) value on the reward of the other person (Shu, Kryven, Ullman, & Tenenbaum, 2020; Ullman et al., 2009; Wu et al., 2023).

Here, we draw on this framework to build a simple computational model of praise. We start with the assumption that people act so as to maximize their expected utility U:

$$U(s, a) = R(s) - C(a), \tag{2}$$

where $R(s)$ is the reward R of being in state s, and $C(a)$ is the cost C of taking action a. Having observed someone's action a allows us to infer how much reward R they must have placed on achieving state s. An agent would only take an action if its cost $C(a)$ is less than the reward $R(s)$. So, knowing the cost of an action $C(a)$ places a lower bound on $R(s)$.

In our setting, we're interested in the reward that an agent places on benefiting another agent. To compute this reward, we need to consider what costs the agent would have incurred in the relevant counterfactual scenario $C(a')$ in which the agent had just pursued a selfish goal. In our setting, a natural value for $C(a')$ is the cost that it would have taken for the agent to go straight to the office without picking up a coffee for their co-worker. So, we can compute a lower bound on the reward that the agent places on a

given state in the following way:

$$R(s) > \sum_{t=0}^{T} C(a_t) - \sum_{t=0}^{T'} C(a'_t). \tag{3}$$

where $C(a)$ is the actual cost, and $C(a')$ the counterfactual cost. Note that we added an additional detail here: a time index t. This allows us to consider settings in which an agent takes multiple actions. The overall cost is then simply the sum of the cost associated with each individual action. In our setting, we equate costs with effort. In the experiments described above, participants were informed that the cost of traversing a blue tile was 1, and traversing a red tile was 3 (red tiles indicate busier areas, so it takes the agent more effort to go through these).

For example, consider the scenario shown in Figure 3a. In this case, the actual cost that the agent incurred was 23. Twenty steps on blue tiles and one step on a red tile. If the agent had gone directly to the office without picking up the coffee, the counterfactual cost of taking that path would have been 13. So, for this example, the action cost of picking up the coffee and going to the office was 23 (compared with doing nothing), and the additional action cost was 10 (compared to what the cost would have been if the agent had gone to the office directly). We assume that once the agent is located on the square with the coffee shop, picking up the coffee doesn't incur any additional cost.

We equate the reward that the agent achieves with their desire to bring their co-worker a coffee. We predict that the greater this desire is, the more deserving of praise they are (for highly valuing their co-worker's benefit). So, the extent to which an agent receives praise is predicted to be:

$$\text{praise} = R(s). \tag{4}$$

In words, an agent is praised based on the reward R that an observer infers the agent places on bringing about state s (which is lower-bounded by the cost they were willing to incur, as shown in Equation 3).

Different models for the cost of action

We implement several versions of the model that differ in how the agent's desire to bring their co-worker a coffee is calculated. In the *action cost model*, we set $C(a'_t)$ in Equation 3 to 0. This model assumes that the agent wouldn't have incurred any cost at all had they not picked up the coffee for their co-worker (they would have just stayed at home). This model predicts that agents will receive more praise the more costly the path was they actually took. In contrast, the *additional action cost model* sets $C(a'_t)$ to the cost that the agent would have incurred had they gone straight to the office without picking up the coffee. This means that if the coffee shop is on the fastest route that the agent would have taken anyhow, the additional action cost would be zero. Finally, we include a *full model* that uses both types of action cost to predict participants' praise judgments. This model assumes that people are weighting both factors when attributing praise: how much effort an agent exerted overall (action cost), and how much more effort they exerted than they would have if they had gone straight to work instead (additional action cost). Prior work has shown that when people attribute responsibility, they care both about how much

effort a person actually exerted, and how much more they could have exerted (Xiang et al., 2023).

Applying the three models to the example in Figure 3a, the action cost model predicts that $R(s) = \beta \cdot 23$, the additional action cost model predicts that $R(s) = \beta \cdot 10$, and the full model predicts that $R(s) = \beta_1 \cdot 23 + \beta_2 \cdot 10$. The action cost model and additional action cost model are special cases of the full model where one of the β parameters is set to 0. The simple models have one free parameter, and the full model has two. We fitted these parameters to the data. Because the full model is more flexible than the other two models, we used cross-validation when comparing model performance to take into account both model complexity and fit.

Inferring effort from praise

With the generative model that maps from the effort that an agent took to the amount of praise they deserve (described above), we can use Bayes' rule to infer how much effort someone must have exerted from the amount of praise that they received. In our experiment, participants had to infer which of two maps that show what path an agent took, was more likely given that an agent had received "low", "medium", or "high" praise (see Figure 3b). We can compute the posterior probability of the agent having taken the path shown in map 1 given a praise judgment $p(m_1|\text{praise})$, as

$$p(m_1|\text{praise}) = \frac{p(\text{praise}|m_1) \cdot p(m_1)}{p(\text{praise}|m_1) \cdot p(m_1) + p(\text{praise}|m_2) \cdot p(m_2)}, \tag{5}$$

where $p(\text{praise}|m_1)$ is the likelihood of a particular praise judgment in map 1 and $p(m_1)$ is the prior probability of map 1. The denominator sums over the two possibilities. We assume a uniform prior over the two different maps (i.e., $p(m_1) = p(m_2) = 0.5$).

To compute the likelihood term in Equation 5, we do the following: across the different maps in our experiments, we rescale the action cost and the additional action cost to each range from 0 to 1. We then compute the likelihood of the different praise terms ("low", "medium", or "high" praise) by centering a separate Gaussian distribution at 0, 0.5, and 1. We fit one parameter σ for the standard deviation in the Gaussian distributions to the data. This allows us to get a graded likelihood for the different possibilities. For example, if the (rescaled) cost for m_1 was 0.3 (rescaled), it's most likely that the agent would receive "medium" praise, somewhat less likely that they would receive "low" praise, and the least likely that they would receive "high" praise.

So, the inference model that uses praise labels (rather than continuous praise values) to infer what happened is

$$p(m_1|\text{praise}_{\text{label}}) = \frac{\mathcal{L}(\text{praise}_{\text{label}}|\text{praise}) \cdot p(\text{praise}|m_1) \cdot p(m_1)}{\sum_{i=1}^{2} \mathcal{L}(\text{praise}_{\text{label}}|\text{praise}) \cdot p(\text{praise}|m_i) \cdot p(m_i)}, \tag{6}$$

where $\mathcal{L}(\text{praise}_{\text{label}}|\text{praise})$ expresses the likelihood of a given $\text{praise}_{\text{label}}$ (e.g., "low" praise), given the predicted praise as computed with Equation 4, and the denominator sums over the two possibilities.

Let us illustrate how the inference part of the model works via applying the additional action cost model to the example shown in Figure 3b. Here, David received "low"

praise for picking up the coffee. In map 1, the additional action cost of the agent's path was 0, whereas in map 2, the additional action cost was 4. Here, picking up the coffee took David 6 steps in total, whereas directly going to the office would have only taken 2 steps. The model then computes the likelihood of receiving "low" praise given an additional action cost of 0 (for map 1) and of 4 (for map 2). Because the Gaussian for "low" praise is centered on 0, the likelihood of having received low praise is higher for map 1 compared to map 2. Accordingly, the model infers that map 1 was more likely.

In this example, we used the additional action cost to compute praise. But we can also use the action cost model instead. The action cost for the agent's path in map 1 is 15 and it's 6 for map 2. Because the action cost is lower in map 2 compared to map 1, this version of the model infers that map 2 is more likely than map 1 given that David received "low" praise. The full model would base its inferences on a weighted combination of the two cost types.

Experiment 2a: Predicting praise from effort

In Experiment 2a, we look at people's judgments about how deserving of praise a person is for picking up a coffee for their co-worker. We hypothesize that the more cost one person incurs for another person's benefit, the more praiseworthy they will be deemed. We compare participants' judgments against the predictions of three versions of our model: one that only considers the actual action cost that the agent incurred, one that only considers the additional action cost, and one that considers both types of action cost.

Methods

The experiment's pre-registration can be found here: `https://osf.io/s9dqw`

Participants

50 participants (*age*: M = 25, SD = 6; *race*: 6 African American/Black, 3 Asian American/Asian, 34 White/Caucasian, 7 other) were recruited through Prolific.[5] Participants received a compensation at a rate of $11 per hour. In order to participate, they had to be at least 18 years old and have an approval rating of at least 90%.

Because we analyze the results by comparing participants' responses to the predictions of different models, there is no straightforward way of performing a power analysis. For experiments that include computational modeling, we chose to run 50 participants per condition which is consistent with prior research that has used similar paradigms to the ones we use here (e.g., Gerstenberg et al., 2018; Kirfel et al., 2022; Langenhoff et al., 2021).

Procedure & Design

In the instructions phase, participants learned about fictional employees going to work. They sometimes go straight from their home to work, and other times they pick up a coffee first. Sometimes the coffee shop is along the way, and other times the employee would need to go out of their way to pick up the coffee. The towns have low-traffic areas (blue grid cells) and high-traffic areas (red grid cells) that take three times as long to traverse.

[5] Due to an experiment error, we did not record gender information for Experiment 2.

Before proceeding to the experiment, participants had to complete two comprehension questions that established (1) that they could identify which of two paths would take longer, and (2) that they understood that traveling through a red cell takes three times longer than traveling through a blue cell. If participants answered either of the questions incorrectly, they were re-routed to reading the instructions again. After correctly answering both questions, participants proceeded to the main experiment.

Participants viewed 48 trials in random order. Figure 3a shows an example trial (see Table B1 for the full list of trials). The trials were selected to vary the action cost (between 6 and 25) and additional action cost (between 0 and 22). We also made sure that the two types of costs were not too highly correlated across the different trials so that we could tease apart how much each type of cost affects participants' praise judgments. Across the trials, the correlation of the two types of cost was $r = 0.60$.

Each trial consisted of a 7×7 grid with three locations: the employee's home (starting point), the coffee shop, and the office (end point). Overlaid on the grid was a path that goes from home via the coffee shop to the office. Between trials, we manipulated the locations of the home, coffee shop, and office, as well as the red regions of high traffic. On each trial participants were asked "How much praise does Person 1 deserve for picking up a coffee for Person 2?" (see Figure 3a). They responded on a slider from "none" (0) to "a lot" (100). The slider handle appeared upon clicking on the slider track. The names of "Person 1" and "Person 2" in the query were randomly selected without replacement from the top 50 male and female baby names over the last hundred years in the United States. At the end of the experiment, participants were asked for demographic information and for feedback on the experiment in a open text form. It took participants 9.4 minutes on average (SD = 4.6) to complete the experiment.

Results

Figure 4 shows the results of four scenarios together with the predictions of three different models: the *action cost model*, the *additional action cost model*, and the *full model* which predicts praise judgments by a weighted linear combination of both action cost types. In scenario 1, all three models are closely aligned with participants' judgments. Praise judgments are high here because the action cost was high, and the path to the office without picking up the coffee would have had a much lower action cost. In scenario 2, praise judgments are lower on average, and the *action cost model* predicted even lower ratings. The person only travels a relatively short path but it would have been even shorter without picking up the coffee. In scenario 3, participants' praise judgments are even lower. Here, the *action cost model* predicts a higher praise judgment because the person travels a long path. However, the coffee shop was along the way to work, so the *additional action cost model* correctly predicts the low praise judgment. Finally, in scenario 4, praise judgments are low again. This time, the *additional action cost model* predicts a slightly higher praise judgment because the coffee shop was out of the way.

Figure 5 shows how well the three models capture participants' judgments across all scenarios. We performed approximate leave-one-out cross-validation to compare model performances against one another.[6] Table 2 shows the result of this analysis. We fitted

[6] Cross-validation is a statistical technique to evaluate a model's performance while controlling for over-

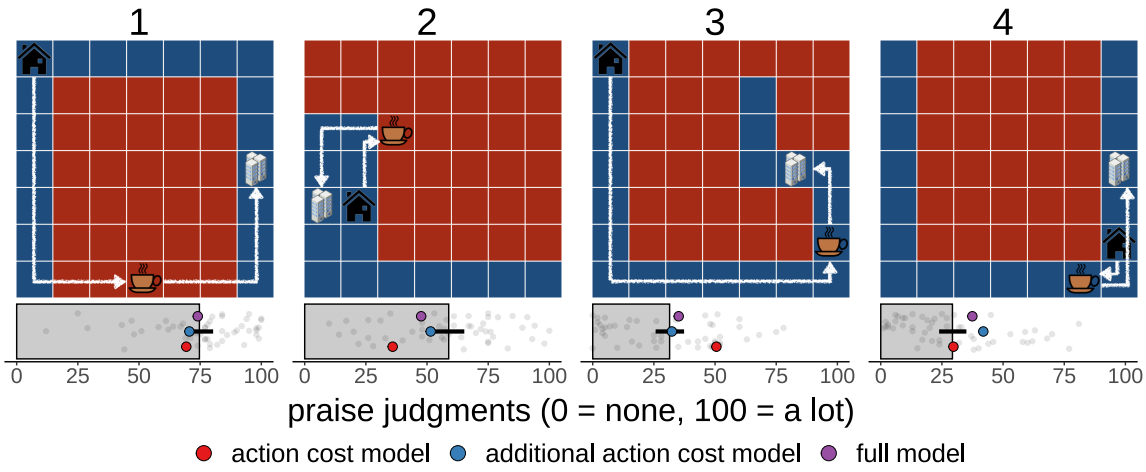


Figure 4

Experiment 2a – Participant judgments and model predictions for four scenarios. Note*: The bars show mean praise judgments with 95% bootstrapped confidence intervals. Gray points show individual participant judgments. Colored points show model predictions.*

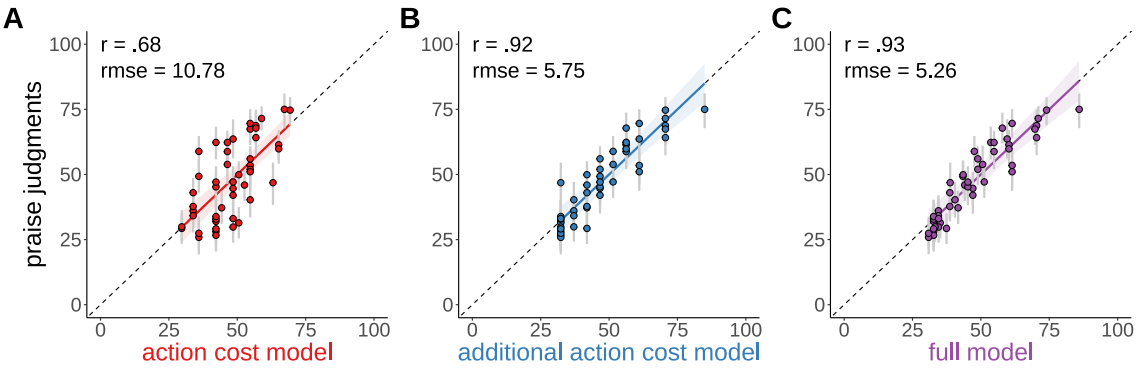


Figure 5

Experiment 2a – Relationship between the predictions of three models and participants' mean praise judgments across the 48 scenarios. Note*: Error bars on the data points are 95% bootstrapped confidence intervals. The ribbon for each regression line shows the 95% credible interval.*

fitting. In cross-validation, the data is split into a training set and a test set. The model's free parameters are first fit to the data in the training set, and the model's predictions are then evaluated against the data in the test set. This procedure is performed multiple times for different training-test splits. More flexible models generally fit the data in the training set better than less flexible models. However, this does not mean that they generalize better to the test set, because it's possible that they overfit to noise in the training set. In this way, cross-validation provides a way to penalize overly flexible models. The goal of this analysis is to establish which model best trades off complexity and fit to the data. Cross-validations differ in how they perform the training-test splits. In leave-one-out cross-validation, all of the data except for one data

four models to participants' judgments. A *baseline model* that only included an intercept as predictor, the *action cost model*, the *additional action cost model*, and the *full model* which uses both predictors. We fitted each of these models as Bayesian linear mixed effects models with random intercepts and slopes.

The results of the cross-validation show that the *full model* captures participants' praise judgments best overall (see Table 2). However, in terms of correlation and root mean squared error, the *full model* only performs slightly better than the *additional action cost model*. The *action cost model* performs clearly worse. We also fitted each model to individual participant judgments, and again used cross-validation to determine best performance. Here, we found that the majority of individual participants are best explained by the *additional action cost model* ($N = 31$), more than by the *full model* ($N = 13$).

Discussion

How much praise does a person deserve for picking up a coffee for their co-worker? In our experiment we found that a good predictor is how much additional effort they incurred to do the favor. People seem to compare how much additional effort a person expended compared to how much effort it would have taken them without doing the favor (Richens, Beard, & Thompson, 2022; Wu et al., 2023). In our setting, this means concretely that people compared the person's actual path against a counterfactual path of directly going

Table 2

Experiment 2a*. Model comparison. The 'model' column shows what predictors were included in each model. The baseline model just includes an intercept as a predictor whereas other models include additional predictors. The 'intercept', 'action cost', and 'additional action cost' columns show the mean of the posterior distribution for this parameter together with the 95% credible interval. 'r' and 'rmse' show the Pearson's correlation and root mean square error. 'Δ elpd' shows the difference in expected log predictive density using approximate leave-one-out cross-validation between the best-fitting model (indicated by 0) and the other models together with the standard error in parenthesis. Lower numbers indicate worse performance. '# best fit' shows the number of participants whose judgments were best captured by each model (as determined via crossvalidation).*

model	intercept	action cost	additional action cost	r	rmse	Δ elpd	# best fit
baseline	47.54 [43.64, 50.97]			0.00	14.65	−786.71 (40.16)	1
action cost model	17.19 [9.35, 25.22]	2.09 [1.69, 2.5]		0.68	10.78	−480.64 (35.25)	5
additional action cost model	32.34 [27.04, 37.65]		2.39 [1.91, 2.85]	0.92	5.75	−53.14 (15.46)	31
full model	25.44 [18.58, 32.73]	0.61 [0.29, 0.94]	2.09 [1.58, 2.6]	0.93	5.26	0 (0)	13

point are in the training set, and the remaining data point is in the test set. This split is then performed as many times as there are data points. However, because fitting the parameters in Bayesian models can take a long time, statisticians have developed faster techniques that approximate the results of this analysis (Vehtari, Gelman, & Gabry, 2017). A measure of the relative model performance is the difference in expected log predictive density 'Δ elpd'. Intuitively, this measure captures how much two models differ in how well they predict the held-out data in the test set. For the best model out of the models that were considered Δ elpd = 0 and for inferior models the value is negative. Following a standard rule of thumb in the literature (see Vehtari et al., 2017, for details), we treat two models as differing credibly in their performance if the difference in the expected log predictive density (Δ elpd) is twice as large as the standard error of that estimate.

to work without picking up a coffee. According to the model we proposed, this additional effort (or cost) translates into praise because it demonstrates how strong the person's desire was to do their co-worker a favor. Going further out of one's way to get a coffee for someone else yielded higher praise than simply picking up a coffee along the way.

While most of the variance in people's praise judgments was explained by the *additional action cost* predictor, participants also cared to some extent about the *actual action cost*. This result is in line with recent work finding that both actual effort and counterfactual effort influence responsibility judgments (Xiang et al., 2023). While the model that incorporated both action cost and additional action cost best fitted participants overall, the majority of individual participants were best fitted by the model that only used additional action cost as a predictor.

Experiment 2b: Inferring effort from praise

Experiment 2a demonstrated that people systematically assign praise as a function of the effort that someone put in for another person's benefit. Experiment 2b investigates whether people can use praise judgments to draw inferences about what happened. If people understand that high praise is assigned to those who go out of their way to help someone else, they should be able to infer that a person who received high praise must have incurred a high cost.

Methods

The experiment's pre-registration can be found here: `https://osf.io/yz2nk`

Participants

50 participants (*age*: M = 28, SD = 10; *race*: 14 African American/Black, 2 Asian American/Asian, 28 White/Caucasian, 6 other) were recruited through Prolific. Participants were reimbursed at a rate of $11 per hour. In order to participate, they had to be at least 18 years old and have an approval rating of at least 90%.

Procedure & Design

The experiment instructions were largely identical to those of Experiment 2a. After having passed the comprehension check questions, participants first assigned praise themselves on three trials. This time, rather than providing a continuous praise judgment, participants were asked "How much praise does Person 1 deserve for picking up a coffee for Person 2?", and selected among three options ("low", "medium", "high"). These trials consisted of situations where the person picking up coffee incurred a lot of additional cost (25 units), some additional cost (6 units), or none (0 units).

After completing the praise judgment trials, participants moved on to the main phase of the experiment. They were instructed that they would be told how much praise a person received for getting coffee, and that their task was to figure out which of two options had happened. Figure 3 shows an example inference trial (see Table B2 for the full list of trials). The 36 inference trials consisted of two 7×7 grids as described above, labeled as "Path 1" and "Path 2". They were told "Person 1 gave Person 2 [low, medium, high] praise for picking up the coffee. Which path do you think Person 2 took?" They responded on a

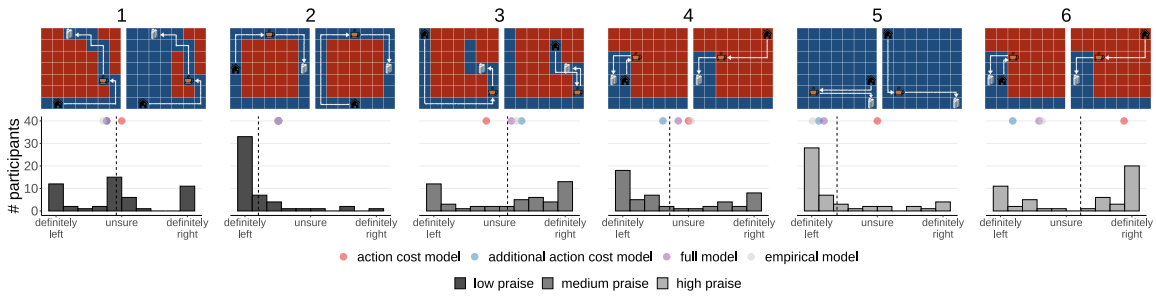


Figure 6

Experiment 2b – Participant inferences about which map was more likely given a particular praise judgment for a selection of scenarios. Note*: Bars show binned counts of participant responses. The dashed vertical line shows the mean response. The colored dots at the top show the model predictions. The shading of the bars indicates how much praise the agent had received (either "low", "medium", or "high").*

slider ranging from "definitely Path 1" (0) to "definitely Path 2" (100), with the midpoint labeled "unsure". On average, it took participants 18 minutes (SD = 9.9) to complete the experiment.

Results

Figure 6 shows participants' inferences for a selection of scenarios. In scenarios 1 and 2, participants learned that the person had received *low praise*. In scenario 1, participants were overall unsure about which of the two maps showed what actually happened. There was a tri-modal response distribution with some participants clearly favoring the left map, some the right map, and some being unsure. Because the path that the person took was identical in both situations, the *action cost model* predicts a rating right in the middle. However, because in the map on the right, an alternative path to the office with lower action cost would have been available, the other models predict that the left map is somewhat more likely given that the person received low praise. In scenario 2, participants were quite certain that it was the person on the left who had received low praise. In the map on the right, a much shorter path would have been available if the person had wanted to go to the office without picking up the coffee. All models capture participants' inferences on this scenario although they underestimate how strong this inference was.

In scenarios 3 and 4, the person had received *medium praise*. In scenario 3, participants were overall relatively unsure what happened but have a slight preference for the right map. This preference goes against the predictions of the *action cost model*. For that model, the likelihood of receiving a medium praise judgment is greater for the left map with the longer path, compared to the right map with the shorter path. In scenario 4, participants inferred that the map on the left was more likely. In the map on the right, the person didn't incur any additional cost.

In scenarios 5 and 6, the person had received *high praise*. In scenario 5, participants inferred the map on the left because the person went out of their way to grab the coffee, whereas for the map on the right, the coffee was on the way to the office. The *action cost*

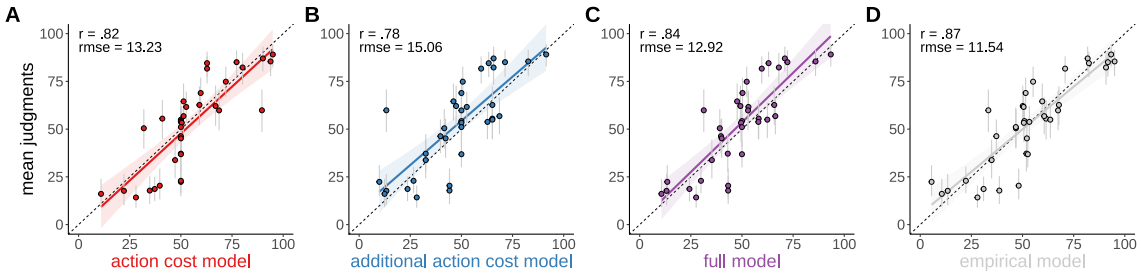


Figure 7

Experiment 2b – Scatter plots showing the relationship between model predictions and participants' mean judgments across scenarios. Note*: Error bars on the data points show 95% bootstrapped confidence intervals. The regression line ribbons show 95% confidence intervals.*

model predicts that people should be uncertain in this case because the action costs were the same in both situations. In scenario 6, participants judged that the map on the right was more likely given that the person had received high praise. This inference goes against the prediction of the *additional action cost model*. In the map on the left, the person incurred a relatively high additional cost by picking up the coffee compared to how easily they could have gotten to the office otherwise. In the map on the right, the coffee shop was along the way to the office.

Figure 7 shows how well the different models capture participants' inferences across all 36 scenarios. To compute the model predictions, we fitted one parameter across all models that captures the standard deviation in the Gaussian likelihood functions that maps from continuous praise judgments to the three different praise terms (see Equation 6). The best-fitting parameter was $\sigma = 0.4$. Otherwise, each model's predictions are based on the parameters that were fitted to participants' judgments in Experiment 2a.[7]

The *empirical model* that uses participants' prediction judgments from Experiment 2a best explains their inferences in this experiment. The *full model* which uses both the action cost and additional action cost as predictors also captured participants' inferences well. Note that it wasn't guaranteed that this model would fit better than the models which only use action cost or additional action cost as predictors in this experiment, because we didn't fit any new parameters here that would otherwise have benefited a more flexible model with more parameters. Surprisingly, the *action cost model* captured participants' inferences slightly better than the *additional action cost model*, even though the latter had been a better account of participants' praise judgments in Experiment 2a.

Discussion

Participants made systematic inferences about what happened from knowing how much praise a person received. The results of Experiment 2b show that we can capture

[7]This means that for this experiment, all three models share the same free parameter σ that we fitted to the data. Because the models don't differ in the number of free parameters, we can compare their performance directly (without the need for cross-validation).

these inferences based on the predictions they made in Experiment 2a. For example, when people hear that a person received "high" praise, they infer that it must have been the person who incurred a greater (additional) cost to get the coffee. So, people can invert their generative model of how action costs translate into judgments of praise, to infer action costs from praise.

Interestingly, while this general relationship between predictions and inferences held up, there was also some degree of mismatch. When participants judged praise from costs (Experiment 2a), the *additional action cost* model captured their judgments well. In contrast, when participants made inferences about costs from praise (Experiment 2b), these inferences were more strongly influenced by the actual action cost (rather than the additional action cost). For example, in scenario 6 in Figure 6, participants inferred that the person who had received "high praise" was more likely to have been the one on the right, even though for that person, the coffee shop was along the way to the office. The person on the left had incurred a lower action cost overall, but had gone out of the way (relatively speaking) to get the coffee.

What could explain the mismatch between praise predictions and inferences based on praise? The inference condition is cognitively more demanding than the prediction condition. If participants carry out the same steps as our computational model, then they have to compute for each scenario how likely the given praise would be, and then make an inference based on that relative likelihood. It is possible that some participants relied on a simpler strategy in their inferences, and just compared the actual path lengths in both maps (without considering what the alternative route for each scenario would have been if the person had directly gone to the office). Scenario 1 in Figure 6 shows a pair of situations where the person's path was identical. A relatively large number of participants were unsure about what happened based on having heard that person received "low praise" (even though the person on the right incurred significant additional costs).

Our generative model assumes a monotonic relationship between effort and praise: the more effort the co-worker exerted, the stronger the model infers the co-workers desire must have been to do a favor. However, it is plausible that a strictly monotonic relationship may not hold in all circumstances. For example, if a co-worker consistently helps out in a supererogatory way , one may infer that they have an ulterior motive. Or, one may end up feeling uncomfortable with that much attention and the possible expectation of having to return the favor. Given that an expression of praise is a strong signal for encouragement, one could imagine a non-monotonic function from effort to praise, where low and medium amounts of effort generate monotonically proportional praise but excessive effort results in lower praise. While we don't find evidence for a non-monotonic relationship in our paradigm, it's plausible that it could arise in other settings, in which case our generative model could be adapted accordingly.[8]

Experiment 3: Inference from blame when cooperation breaks down

In Experiments 1 and 2, participants had to infer what happened based on one person having blamed (Experiment 1) or praised (Experiment 2) another. In Experiment 3, we focus on people's inferences from social evaluations when a group failed to optimally coor-

[8] We thank an anonymous reviewer for raising this point.

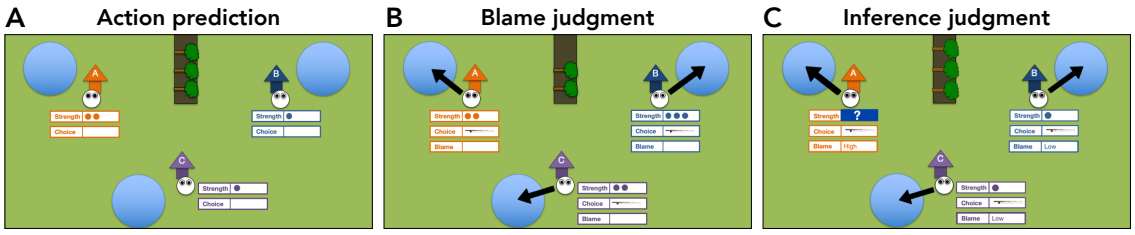


Figure 8

The fishermen paradigm. *Each fisherman has a strength which determines how many fish they can catch or how many trees they can remove from the blocked road. Fishermen know each other's strength but have to decide without communicating whether to fish or remove trees. The fishermen can only sell their fish if the road is unblocked.* ***A*** *In Experiment 3a, we ask participants what action a fisherman should take.* ***B*** *In Experiment 3b, we ask participants how much blame a fisherman deserves for a suboptimal group outcome.* ***C*** *In Experiment 3c, we ask participants to 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.*

dinate 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. This setup presents a number of novel challenges. When agents act cooperatively as part of a group, their actions also reflect how they think others will act. Together, they need to solve a complex coordination problem where different agents often take different actions in the hope that, if all goes according to plan, the expected benefit of the group will be maximized.

To capture this, we developed a more complex model of blame attribution that takes into account what an agent should have done in the context of a group decision, and what causal role their action played for bringing about the suboptimal group outcome. We then see what inferences participants can draw from blame judgments whereby, this time, there might be several pieces of missing information that participants are asked to infer. Overall, this setup presents a more challenging test bed, both for modeling blame judgments, and for studying what inferences people can make based on these judgments.

Overview of the experimental paradigm

Participants in our experiment were introduced to a village of fishermen (see Figure 8). 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, or how many trees they can remove from the road. Going fishing or clearing 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. However, it's common knowledge how strong everyone is, and how many trees block the road. At the end of the day, if the road has been cleared, the fishermen equally distribute the money earned from

the fish sacks they have collected. If the road isn't cleared, they receive nothing. So, the fishermen's incentives are fully aligned: they want to clear the trees and catch as many fish as possible.

Consider the situation shown in Figure 8a. Fisherman A has strength 2, 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 one sack of fish. For that to happen, either fisherman B or C should go fishing, and the other two should clear the trees. Now consider what happened in Figure 8b. Here, two trees are blocking the road, fishermen A and C have strength two, and fisherman B has strength 3. All three fishermen ended up going fishing. To what extent should each of the fishermen be blamed for the suboptimal outcome? Finally, consider the situation in Figure 8c. Given that fisherman A was blamed highly for the suboptimal outcome and fishermen B and C received low blame, how strong do you think fisherman A was?

The three examples in Figure 8 illustrate the three studies in Experiment 3. In Experiment 3a, we asked participants to judge what action a fisherman should take. In Experiment 3b, participants judged the extent to which a fisherman was to blame for a suboptimal group outcome. And in Experiment 3c, participants' had to infer hidden aspects of the situation based on information about how much each fisherman was blamed. Before discussing the experiment results, we describe our computational model.

A computational model of blame for cooperation failures

We describe the action model, blame model, and inference model in turn.

Action model

How should an agent choose what action to take (see Figure 8a)? In a purely cooperative coordination game, individuals should attempt finding an optimal strategy to maximize the group's expected reward (Schelling, 1980). If there is only one way for the group to succeed, and all group members are rational, finding the solution is simple. However, when there is more than one way for the group to get the optimal reward, and these have conflicting strategies for each agent, the choice is less clear. Here, we compare two different models of how agents choose their actions: a *recursive reasoning model*, and a *heuristic choice model*.

Recursive reasoning model

In the fishermen paradigm, each fisherman has to choose whether to go fish or clear trees. According to the *recursive reasoning model*, agents choose actions to best respond to their companions at a level k depth of reasoning (Camerer, Ho, & Chong, 2004; Costa-Gomes, Crawford, & Broseta, 2001; Stahl II & Wilson, 1994; Yoshida, Dolan, & Friston, 2008). At level $k = 0$, each agent randchooses one of two actions. 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 others' random actions. At further levels 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 that are associated with the two actions that each fisherman can take

$$p^k(a_f) = \frac{\exp(\beta_r \hat{r}_k[a_f])}{\sum_{a_f \in \text{actions}} \exp(\beta_r \hat{r}_k[a_f])}, \tag{7}$$

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 the highest expected reward; Luce, 1959).

Heuristic choice model

According to the *heuristic choice model*, agents first figure out what combinations of actions would lead to an optimal reward, and then choose uniformly between those actions. For example, in the situation shown in Figure 8a, the model predicts a 50% probability for fisherman B (or C) to clear the trees and a 100% probability for fisherman A to clear the tress. Here, there are two sets of actions leading to optimal reward: fisherman A and B clear the trees and fisherman C goes fishing, or fisherman A and C clear the trees and both fisherman B goes fishing. In both situations, fisherman A clears the trees, so his action is clear. However, for fisherman B (or C) there is one scenario in which he should fish, and one in which he should clear the trees, so this model predicts that he will choose either action with 50% probability, such that $p(a_f) = 0.5$.

Blame model

How much blame does each agent deserve for a suboptimal group outcome (see Figure 8b)? As we mentioned in the introduction, prior work has shown that people care about two main aspects when attributing blame (Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa et al., 2021; Wu et al., 2023): first, they care about what an agent should do and what their action reveals about them. Second, they care about what causal role an agent's action played. We develop a blame model for the fisherman domain that captures both aspects. We model what the action reveals about the person by considering the extent to which the agent did what they should have done. If an agent deviated from what they should have done, this reveals either that they didn't care or that they didn't sufficiently think things through (both of which are blameworthy). We call this the *rationality* component of the model. We model an agent's causal role by considering whether their action was pivotal for the suboptimal outcome and, if not, how close it was to having been pivotal. We call this the *pivotality* component of the model.

Blame for suboptimal actions: What should the agent do?

This component of the model considers how optimal an agent's action was (Johnson & Rips, 2015). The more clear it was what an agent should have done, the more they get blamed for not doing that. The action models (recursive reasoning or heuristic choice) dictate what action the agent should take, $p(a_f)$. So the (ir)rationality component of blame is simply the inverse of that probability:

$$\text{blame}_{\text{rationality}} = 1 - p(a_f). \tag{8}$$

An agent receives blame for a suboptimal group outcome to the extent that they didn't take the rational action.

Blame for pivotal actions: How much did the agent's action matter?

In Experiments 1 and 2, the agents' actions were always pivotal. This means that the outcome would have been different, had they acted differently. For example, in Experiment 2, the person wouldn't have received their coffee, if their co-worker hadn't brought it. Once multiple agents contribute to an outcome, it's possible that their individual action ends up not mattering in the actual situation. In elections, for example, the winner is often overdetermined, and it's rare for an individual vote to be pivotal (such that the other candidate would have won, had they voted differently). Chockler and Halpern (2004) developed a model that predicts an agent's responsibility as a function of how close their action was to making a difference to the outcome. Prior work has demonstrated that this model accounts well for how people assign responsibility to individuals for group outcomes (Engl, 2022; Gerstenberg & Lagnado, 2010; Gerstenberg, Lagnado, et al., 2023; Lagnado & Gerstenberg, 2015; Lagnado et al., 2013; Langenhoff et al., 2021; Sosa et al., 2021; Zultan et al., 2012). Here, we use this model to capture the extent to which an agent's action mattered for a suboptimal group outcome. Formally, the extent to which fisherman f's action was pivotal for the suboptimal outcome is defined as

$$\text{blame}_{\text{pivotality}} = \frac{1}{N_f + 1}, \tag{9}$$

where N_f is the minimal number of other fishermen whose action would have needed to be different such that changing the action of fisherman f would have led to an optimal outcome. In Figure 8b, the pivotality of fishermen A and C is 1 because if either of them had acted differently, the group would have achieved the best possible outcome (i.e., the number of changes to be pivotal is $N_f = 0$ for each of them). 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 ($N_f = 1$). Neither fisherman A nor C are pivotal in the actual situation, but would have been pivotal if fisherman B had gone fishing instead of going for the trees.

Blame for irrational, pivotal actions

We assume that participants blame others by taking into account both the rationality of their action and how much what they did mattered. Concretely, we predict that people's blame judgments are weighted mixture of the blame values predicted based on rationality (Equation 7) and pivotality (Equation 9)

$$\text{blame} = \beta_1 \cdot \text{blame}_{\text{rationality}} + \beta_2 \cdot \text{blame}_{\text{pivotality}}, \tag{10}$$

where β_1 and β_2 determine how much weight is placed on the rationality and pivotality component when assigning blame.

We will compare three models of blame to participants' judgments. The *rationality model* only considers the rationality component. The *pivotality model* only considers the pivotality component. The *full model* considers both components as specified in Equation 10.

Inference model

What can people infer about the situation from knowing how much blame each agent received (see Figure 8c)? In Experiment 3c, we withhold information of at least one situational aspect. For example, we hide how many trees blocked the road, a fisherman's strength, or their action. In Figure 8c fisherman A's strength is hidden. Fisherman A received high blame and fisherman B and fisherman C both low blame for the suboptimal group outcome. What can we infer from these blame judgments about fisherman A's strength?

Just like in Experiment 2, we again use Bayes' rule to invert the generative model to make inferences from social evaluations. This time, we compute the probability of a situation s_i, given an assignment of blame to each fisherman $p(s_i|\text{blame}_{\text{label}})$:

$$p(s_i|\text{blame}_{\text{label}}) = \frac{\prod_{f \in \text{fishermen}} \mathcal{L}(\text{blame}_{\text{label}_f}|\text{blame}_f)p(\text{blame}_f|s_i)p(s_i)}{\sum_{i \in \text{situations}} \prod_{f \in \text{fishermen}} \mathcal{L}(\text{blame}_{\text{label}_f}|\text{blame}_f)p(\text{blame}_f|s_i)p(s_i)}, \quad (11)$$

where s_i is a potential situation (e.g., that fisherman A has strength 2), blame_f is the predicted blame for fisherman f, and $\text{blame}_{\text{label}_f}$ is the blame label that the agent actually received. We assume a uniform prior over possible situations $p(s_i)$.[9]

The model's blame values for each fisherman $p(\text{blame}_f|s_i)$ are deterministic given the parameters of the model. The likelihood function $\mathcal{L}(\text{blame}_{\text{label}_f}|\text{blame}_f)$ involves a comparison of the observed blame labels $\text{blame}_{\text{label}_f}$ against the blame values predicted by the blame model in each situation. In our experiment, fishermen either received "low", "medium", or "high" blame (see Figure 8c). Because the model's predictions are continuous, we converted each qualitative blame label into a value (0, .5, 1, respectively), and computed the likelihood of a blame label given a blame value ($\mathcal{L}(\text{blame}_{\text{label}_f}|\text{blame}_f)$) as normally distributed with standard deviation σ.

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 8c). 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:

$$p(o_j) = \frac{\exp(\beta_d \cdot p(o_j|s_i))}{\sum_{o_j \in \text{options}} \exp(\beta_d \cdot p(o_j|s_i))} \quad (12)$$

The model has two free parameters: σ in Equation 11 for the mapping from continuous blame values (which range between 0 and 1) to discrete blame labels ("low", "medium",

[9] We exclude situations with optimal group outcomes, as participants in our experiment were told that the fishermen were blamed for failing to achieve the best outcome (and it wouldn't make sense to give someone blame for achieving the best outcome).

or "high"), and the temperature parameter β_d in the softmax decision function in Equation 12 for the mapping from continuous inferences over the different options to the discrete choices.[10] We fitted both parameters to the data, and shared the same two parameters across the three versions of the blame model (i.e., the rationality model, pivotality model, and full model). The best fitting parameters were $\sigma = 0.3$ and $\beta_d = 1.2$.

Now that we have a model that predicts fishermen choices, how much blame they receive, and what inferences one can draw from blame judgments, we test the model's predictions in three separate experiments.

Experiment 3a: What should the fishermen do?

In Experiment 3a, we ask participants to judge what fisherman A should do based on information about how many trees are blocking the road, and how strong each fishermen is (see Figure 8a).

Methods

Participants

50 participants were recruited through Amazon Mechanical Turk. We didn't collect demographic information in Experiments 3a and 3b.

Procedure & Design

We asked participants to judge which action fisherman A should take on a sliding scale from "Definitely fish" (0) to "Definitely clear road" (100). They were given a tutorial explaining the fishermen's situation, and asked to answer some comprehension checking questions. We generated different scenarios by considering all unique permutations of 1–3 trees and three fishermen with strengths 1–3, leading to 54 different scenarios (see C1 for the full list of scenarios). Participants were then shown a randomly selected subset of 27 of these. 50 participants were recruited through Amazon Mechanical Turk. This yielded between 23 and 27 observations for each scenario.

Results

Figure 9 shows participants' judgments and model predictions for a selection of scenarios. Is scenario 1, there was one tree blocking the road and the fishermen's strengths were 3, 1, and 2. Participants' were asked to judge what fisherman A (who had strength 3) should do. In this case, participants judged that fisherman A should definitely fish, as did both the recursive reasoning model and the heuristic choice model. In scenario 2, participants still strongly favored "fish" but a little less than in scenario 1. This is captured by the recursive reasoning model. Because both fishermen B and C have strength 1, it's not clear which of them should go for the trees. In contrast, the heuristic choice model predicts that it's clear what fisherman A should do because in all situations that lead to the optimal outcome (selling 3 fish), fisherman A goes fishing. In scenario 3, the predictions between the two models come even further apart. The recursive reasoning model predicts

[10] We didn't need this softmax step in Experiment 2 because in that experiment, participants answered on a continuous slider, whereas here, they need to map from their continuous inferences to discrete choices.

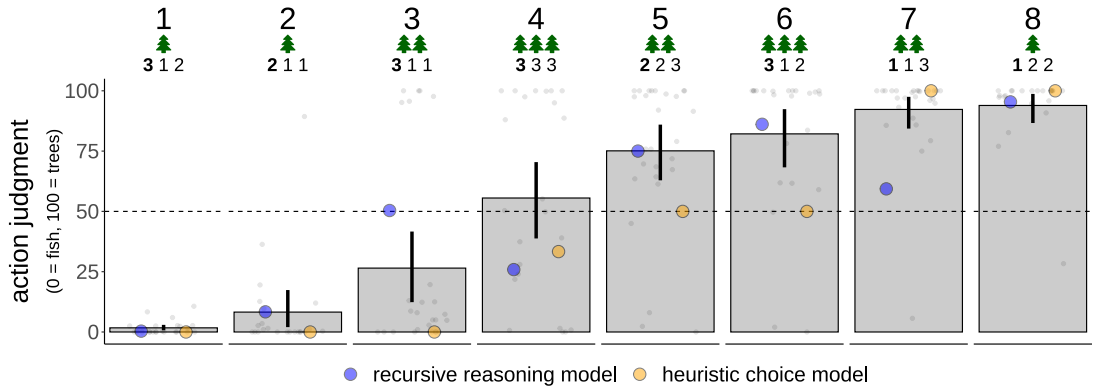


Figure 9

Experiment 3a – Participants' judgments of what action fisherman A should take. Note*: 0 = 'go fishing' and* 100 *= 'clear trees'. The tree symbols indicate the number of trees blocking the road. The numbers underneath indicate the strengths of the three fishermen. For example, in scenario 1, there was one tree blocking the road, and the fishermen's strenghts were 3, 1, and 2. Here, participants judged that fisherman A should go fishing. Bars show mean judgments with 95% bootstrapped confidence intervals. Gray points show individual participant responses (jittered along the x-axis for visibility). Colored points show model predictions. For ease of interpretation, the heuristic choice model's predictions haven't been fitted to the data in this figure.*

that fisherman A should be just as likely to go fishing or go clear trees, whereas the heuristic choice model predicts that he should go fishing (and rely on the other two going for the trees). On average, participants' judgments fall right in between the two models' predictions. In scenario 4, participants are uncertain overall while both models suggest that fisherman A should be more likely to go fish than go for the trees.

In scenario 5, participants' judge on average that he should go for the trees, and the recursive reasoning model captures that. In scenario 6, there are two optimal solutions: either fisherman A clears the trees and the other two fish, or vice versa. Because there are only two optimal solutions and fisherman A's actions are different in each one, the heuristic model predicts 50. The recursive reasoning model captures participants' average judgment that he should go for the trees. Intuitively, going for the trees is better here because if either of the other goes fishing, the group still gets to sell some fish. However, if fisherman A goes fishing, then they won't be able to sell any fish unless both of the others end up going for the trees. While only focusing on optimal solutions would give fisherman A no reason to choose one action over the other in this scenario, by reasoning more deeply about what the others may do, the more robust strategy emerges as the better choice. In scenario 7, the recursive reasoning model doesn't capture participants' judgments. Participants say that fisherman A should go for the trees because both fisherman A and B going for the trees and fisherman C going fishing would result in the optimal outcome. However, the recursive reasoning model assigns some probability to fisherman C going for the trees in which case it would be best for fisherman A to go fish. Finally, in scenario 8, participants judge that fisherman A should definitely clear the tree and both models agree.

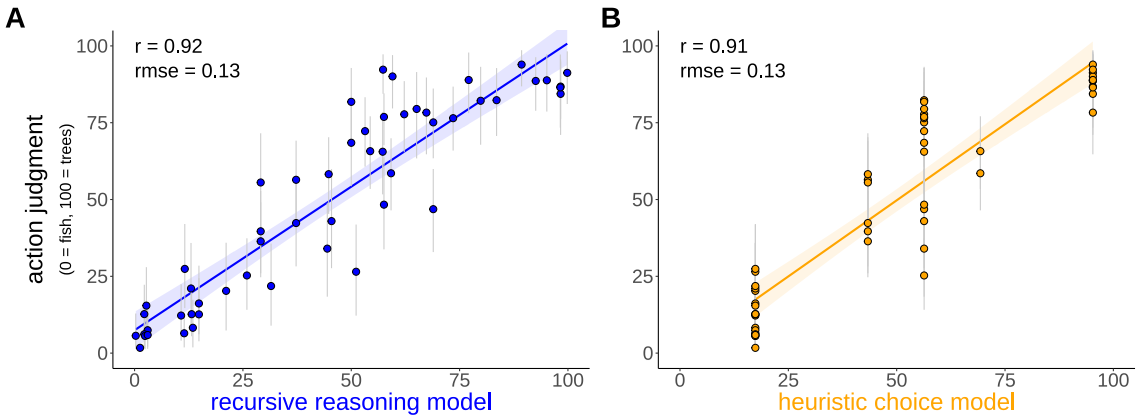


Figure 10

Experiment 3a – Scatter plot showing model predictions and participant mean action judgments. Note*: Error bars show 95% bootstrapped confidence intervals. The regression line ribbons show 95% confidence intervals.*

Figure 10 shows the relationship between model predictions and participants' judgments across all of the scenarios. We fitted the recursive rationality model to participant data using grid search to minimize the squared error between model predictions and data (see Figure 10a). The best-fitting parameters were $k = 3$ and $\beta_r = 1.3$ (see Equation 7).[11] We fitted the heuristic choice model through a linear mixed effects regression (see Figure 10b).

Both models capture much of the variance in participants' judgments. The heuristic choice model only predicts five distinct judgments. This is because there are never more than three optimal solutions in our setting. So the model's predictions (before fitting to the data) are that fisherman A's probability of going to clear trees should be 0%, 33%, 50%, 66%, or 100%. The recursive reasoning model makes more graded predictions that accurately capture participants' judgments in most of the scenarios, but there are also exceptions like scenario 7 shown in Figure 9 (which is the scenario for which the recursive reasoning model shows the largest error).

Discussion

In Experiment 3a, we asked participants to judge what action an agent should take in a cooperative group setting where agents can't communicate with one another. We found that participants' judgments were accurately predicted by a recursive reasoning model that assumes that agents best-respond to the actions that they think the other agents will take. A heuristic choice model that predicts that agents choose actions by solely considering optimal outcomes, also accounted for much of the variance in participants' responses. While the models did a good job of capturing the mean data on the trial level, there was considerable variation between individual participants' judgments. For example, in scenario 3 in Figure 9,

[11] In the grid search, we first defined a range of possible values for each parameter $k = [1, 2, 3, 4]$ and $\beta_r = [1, 1.1, 1.2, ..., 3]$ and then evaluated the model for each possible combination of these parameter values.

some participants judged that the fisherman should fish, whereas others judged that the fisherman should clear the trees. This suggests that participants may have used different strategies in their judgments. In fact, when we fitted both models to individual participant judgments, we found that 26 participants' judgments were better captured by the recursive reasoning model, and 24 by the heuristic choice model.[12]

Experiment 3b: Blame judgment

In Experiment 3b, we ask participants to judge how much blame each fisherman deserves when the group failed to achieve an optimal outcome (see Figure 8b). As Equation 10 shows, we predict that blame judgments are influenced by two factors: the extent to which an agent took the right action (rationality), and how much their action mattered for the group outcome (pivotality).

Methods

Participants

59 participants were recruited through Amazon Mechanical Turk.

Procedure & Design

Participants were first introduced to the fishermen scenario via a tutorial that included a set of comprehension check questions. To continue to the main stage of the experiment, participants had to correctly identify what action a fisherman should take for at least six out of seven scenarios. In the main stage of the experiment, participants judged how much each fisherman was to blame for the group's failure to get the best possible outcome (see Figure 8b). The actions of the fishermen were represented as arrows either towards their pond, or towards the trees on the road. Participants also saw how many fish sacks the fishermen actually collected, as well as the best possible number they could have collected, next to the image. The blame for each fisherman was assessed on a sliding scale from "Not at all" (0) to "Very much" (100).

Since there are many possible combinations of strengths, trees, and choices, we selected a subset of sixty scenarios for which our blame model predicted a range of judgments (see Table C2 for the details of each scenario). Each participant provided judgments for 20–23 of these scenarios, whereby each of the sixty scenarios was judged by at least 16 participants.

Results

Figure 11 shows participants' blame judgments for a selection of cases. In scenario 1, fisherman B received most of the blame. He decided to go fishing when he should have cleared the road. Had he made the correct choice, the group would have been able to sell

[12] For this analysis, we did not fit the two free parameters of the recursive reasoning model to each participant separately but instead used the globally fitted parameters of $k = 3$ and $\beta_r = 1.3$. We then regressed the predictions of this globally fitted model, and the heuristic choice model (which doesn't have any free parameters) to individual participants' responses. We determined which model better captured an individual participant's responses based on the likelihood of the data.

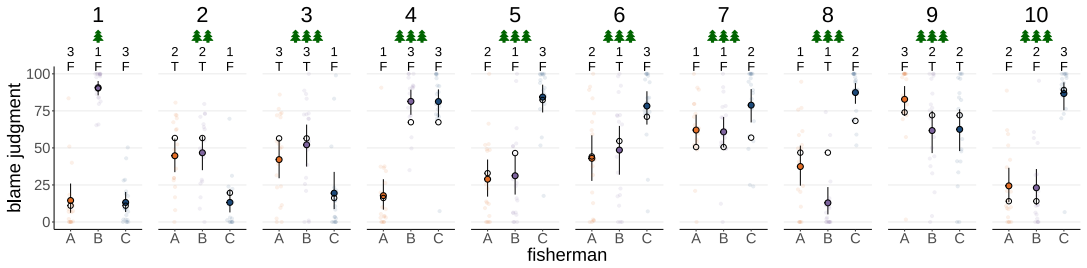


Figure 11

Experiment 3b – Blame judgments and model predictions for a selection of scenarios. Note*: The tree symbols indicate how many trees blocked the road. The numbers below indicate each fisherman's strength. The letters indicate whether a fisherman fished (F) or cleared trees (T). For example, in scenario 1, there was one tree blocking the road, the fishermen's strengths were 3, 1, and 3, and all three fishermen went fishing. Large points show means with 95% bootstrapped confidence intervals. Small points show individual participant responses (jittered along the x-axis for visibility). Unfilled circles show the full model's predictions.*

six sacks of fish. In scenarios 2 and 3, both fisherman A and B share the blame. Here, it wasn't clear who should go clear trees and who should go fishing between the two. In scenario 4, both fisherman B and C went fishing. Notice that the scenario is otherwise the same as scenario 3. The two fishermen with strength 3 received more blame for going fishing than for going to clear the trees and this is predicted by the full model. In scenario 5 and 6, the only thing that differs is fisherman B's action. As predicted by the full model, this affects the extent to which the other two fishermen are blamed. In scenario 5, fisherman C was pivotal. The fishermen would have achieved the optimal outcome had he cleared the trees. In scenario 6, fisherman A is pivotal (because fisherman B went for the trees). As predicted by the model, fisherman A's blame increased in scenario 6 compared to scenario 5, and fisherman C's blame decreased. Scenarios 7 and 8 illustrate a similar effect: fisherman C (who should definitely clear the trees) is blamed more when his action was pivotal than when it wasn't. Scenario 9 is interesting in that each of the fishermen took an action that they shouldn't have taken. In scenario 10, it's clear that most of the blame rests on fisherman C – it was clear that he should have gone for the trees and his action was pivotal.

Figure 12 shows how well the different models capture participants' blame judgments. For the rationality model, we used the recursive reasoning model with the best-fitting parameters from Experiment 3a ($k = 3$, $\beta = 1.3$), as that experiment directly tested people's expectations about how an agent should act. We fitted the *rationality model* (Figure 12a), *pivotality model* (Figure 12b), and *full model* (Figure 12c) to participants' blame judgments using Bayesian linear mixed effects models with random intercepts and slopes. We also fitted a *baseline model* which only included a global intercept and random intercepts as predictors. Table 3 shows the results of comparing the models via cross-validation. The *full model* best captured participants' blame judgments overall, and also best explained most individual participants' judgments (36 out of 59). The *rationality model* and *pivotality model* performed similarly and best explained the judgments of 10 individual participants

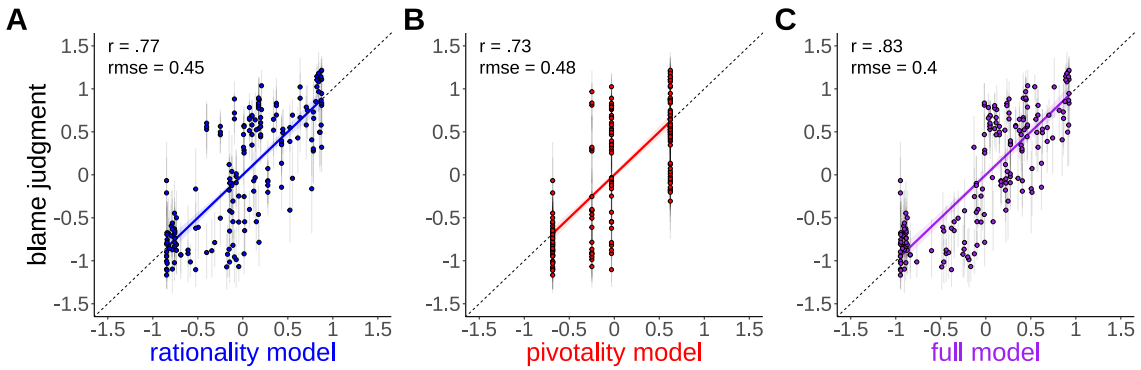


Figure 12

Experiment 3b – Model predictions against participants' z-scored blame judgments. Note*: Error bars on the data points are 95% bootstrapped confidence intervals. The ribbon for each regression line shows the 95% credible interval. Model predictions were fitted to z-scored individual participant judgments.*

each.

Discussion

When participants hold individual agents responsible for a suboptimal group outcome, their judgments are sensitive to whether the agent chose the right action (rationality), and to what extent their action choice mattered (pivotality). In line with prior research (Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa et al., 2021; Wu et al., 2023; Xiang et al., 2023), a blame model that combines both of these components, captures participants' judgments better than lesioned models that rely on either component alone.

Table 3

Experiment 3b. *Model comparison. The 'model' column shows what predictors were included in each model. The baseline model just includes an intercept as a predictor whereas other models include additional predictors. The 'intercept', 'pivotality', and 'rationality' columns show the mean of the posterior distribution for this parameter together with the 95% credible interval. 'r' and 'rmse' show the Pearson's correlation and root mean square error. 'Δ elpd' shows the difference in expected log predictive density using approximate leave-one-out cross-validation between the best-fitting model (indicated by 0) and the other models together with the standard error in parenthesis. Lower numbers indicate worse performance. '# best fit' shows the number of participants whose judgments were best captured by each model (as determined via crossvalidation).*

model	intercept	pivotality	rationality	r	rmse	elpd	# best fit
baseline	0 [−0.03, 0.04]			0.00	0.70	−915.41 (36.43)	3
pivotality	−0.68 [−0.76, −0.61]	1.31 [1.17, 1.45]		0.73	0.48	−304.01 (23.18)	10
rationality	−0.85 [−0.95, −0.75]		1.72 [1.53, 1.94]	0.77	0.45	−139.83 (16.5)	10
full	−0.95 [−1.06, −0.84]	0.69 [0.55, 0.82]	1.19 [0.96, 1.38]	0.83	0.40	0 (0)	36

Experiment 3c: Inference from blame judgments

In Experiment 3c, we ask participants to infer hidden aspects of what happened based on partial information about the scenario, and on information about how much each fisherman was blamed (see Figure 8c).

Methods

The experiment's pre-registration can be found here: `https://osf.io/x37rj`

Participants

50 participants (24 female, 25 male, 1 prefer not say, age: $M = 41$, $SD = 11$) were recruited from Amazon Mechanical Turk and compensated with $2.75. 40 additional participants were excluded because they failed the pre-registered criterion of passing both of the attention checks. In the attention checks, participants were given full information about the scenario and had to correctly say how many fish sacks would be sold.[13]

Procedure & Design

Participants were instructed about the task and required to answer comprehension checks. One question established that they understood that the fishermen shared their earnings equally, and two questions made clear that they knew how many sacks of fish would be sold in each situation. Participants also saw 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 answered the question: "How much is each fisherman to blame for the group's failure get the best possible outcome?". For each fisherman they could select between giving "low", "medium", or "high" blame.

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 8c, 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 scenarios were presented in random order (see Table C3 for the full set of scenarios). It took participants 10.2 ($SD = 5$) minutes on average to complete the experiment.

The scenarios varied what inferences participants were asked to make (trees, choices, and/or strength) and how many pieces of information were missing (from one to three). To

[13] At the time when we conducted the experiment in January 2021, Mechanical Turk had trouble with bots (see Webb & Tangney, 2022). In their response to Webb and Tangney's (2022) article, Keith and McKay (2024) suggest that best practices for running online experiments is to include attention checks, which is what we did for our study. For the studies in Experiment 1 and 2 (which were conducted later), we used Prolific instead of Mechanical Turk.

assess whether participants were sensitive to how much each fisherman was blamed, roughly half of the scenarios involved cases where the situation was held constant but the amount of blame assigned to each agent varied. For example, in scenarios 1 and 2 in Figure 13 there was one fisherman with strength 3 and two others with strength 1, and all went fishing. Whereas in scenario 1, the weak fishermen received high blame and the strong fisherman received low blame, in scenario 2 the pattern was reversed. Similarly, in scenario 4 and scenario 5, 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 scenario 4, whereas one received low blame and the other high blame in scenario 5.

We selected scenarios to make sure that the different model's predictions weren't too highly correlated with one another. For example, in scenario 6 the pivotality model infers that the fisherman B's strength was most likely 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 a strength of 3 was most likely. If his strength had been 2, it would have been reasonable to go fishing (assuming that fisherman C would clear the trees). The reasonableness of fisherman B's going fishing in this case would be incompatible with the high blame he received. Across the 36 scenarios, the correlation between the rationality and pivotality models' predictions was $r = .57$.

Results

Figure 13 shows participants' inferences and model predictions for a selection of eight scenarios. Participants' inferences about what happened are sensitive to how much each fisherman was blamed. For example, scenarios 1 and 2 are identical in terms of the fishermen's strengths and choices (except for shuffling), but differ in how much blame each fisherman received. In scenario 1, both fishermen with strength 1 received high blame and the fisherman with strength 3 low blame. From this information, participants inferred that it was most likely that there was one tree blocking the road. In scenario 2, the fishermen with strength 1 received low blame and the fisherman with strength 3 high blame. Here, participants inferred that 3 trees blocked the road. In scenario 3, the rationality model and the pivotality model make different predictions. All three fishermen went to clear the trees. The two fishermen with strength 2 received high blame, and the fisherman with strength 1 low blame. The rationality model predicts that one tree was blocking the road. The pivotality model predicts that three trees blocked the road. If there had been three trees, then each of the fishermen with strength 2 would have been pivotal for the suboptimal outcome in the actual situation (and the pivotality of the fisherman with strength 1 would be 0 as there are no situations in which him going fishing would lead to the optimal outcome). Participants' inferences in this scenario follow the predictions of the rationality model.

While scenarios 1–3 asked participants to infer the number of trees blocking the street, scenarios 4–6 asked participants to infer one of the fishermen's strength. In scenarios 4 and 5, the strengths of fishermen A and B are the same. Fisherman C takes the same action in both scenarios *and* receives the same amount of blame. However, participants make different inferences about fisherman C's strength based on the actions that the other fishermen took and the blame that they received. In scenario 4, participants think it's most likely that fisherman C's strength was 1 (in which case it would never make sense for him

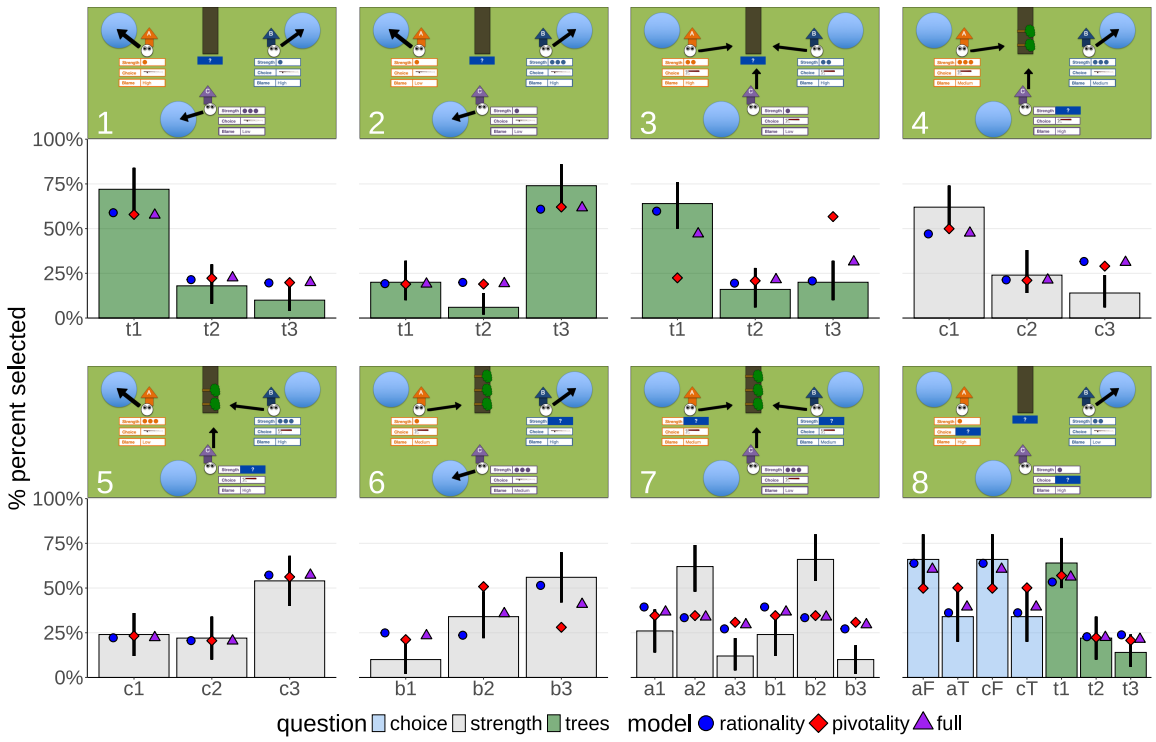


Figure 13

Experiment 3c – 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 full model (purple triangle). Note: For the x-axis labels, t = trees; a, b, c stand for the three fishermen; T = going for the trees, F = going fishing. For example, in scenario 1, participants had to infer how many trees there were. In scenario 4, the missing piece was fisherman C's strength. In scenario 8, participants had to infer fishermen A and C's actions, and how many trees there were.*

to go clear the trees). In scenario 5, participants infer that fisherman C's strength was most likely 3 (it's possible participants realized that fisherman B, who took the same action, also received high blame). Both the rationality model and the pivotality model capture participants' inferences in these two scenarios. In scenario 6, the two models make different predictions. The pivotality model predicts that fisherman B's strength was 2. In that case, fisherman B would have been pivotal for the suboptimal outcome. The rationality model predicts that fisherman B's strength was 3. In that case, it would have been better for fisherman B to go and clear the trees. Participants' inferences aligned more closely with the predictions of the rationality model in this scenario. In scenarios 7 and 8, several aspects of the situation were hidden. In scenario 7, both fisherman A's and B's strengths were unknown. Participants inferred that both of them were most likely to have strength 2. Both of the models predict that their strength could have also been 1. Finally, in scenario 8,

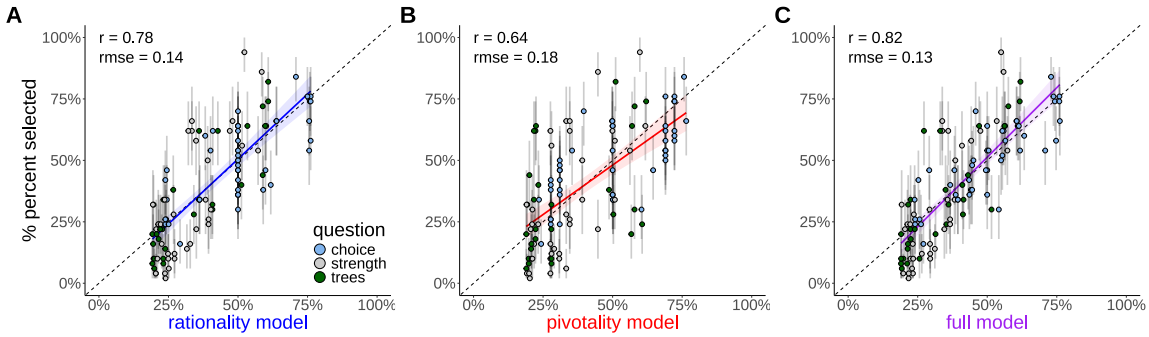


Figure 14

Experiment 3c – Model predictions against participants' inference judgments. *Note: Colors indicate what kind of information participants were asked to infer: A fisherman's choice (blue), their strength (gray), or the number of trees (green). r = Pearson's correlation coefficient, rmse = root mean squared error. Error bars show bootstrapped 95% confidence intervals. The regression line ribbons show 95% confidence intervals.*

three pieces of information were hidden: what fishermen A and C did, and how many trees blocked the road. Participants inferred that both fishermen went fishing, and that it was one tree that blocked the road. If it had been two or three trees blocking the road, then fisherman B would have received high blame for going fishing.

Figure 14 shows how well the different models capture participants' selections across all 36 trials (45 total judgments) in the experiment. These model predictions are based on first having fitted each model to participants' blame judgments in Experiment 3b, and then fitting the standard deviation σ in the function that translates continuous blame judgments to blame labels (see Equation 11), and the temperature parameter in the softmax function that maps from continuous beliefs over possible situations to a discrete choice (see Equation 12). What this means is that, for this experiment, each model has the same number of free parameters, so we can directly compare the model performances against one another. We find that the full model (Figure 14c) which considers both rationality and pivotality captures participants' inferences better than the rationality model (Figure 14a) and the pivotality model (Figure 14b).

Discussion

When group members are blamed for failing to coordinate their actions, people can make systematic inferences about what happened. They infer aspects of the situation and what actions a person must have taken. In Experiment 3b, we had found that participants' blame judgments were well-accounted for by a model that considers how rational a person's action was, and how pivotal it was (i.e., how close it was to having brought about the optimal group outcome). Here, we found that inverting this generative model of blame predicts participants' inferences. Overall, the model does a good job of capturing participants' inferences (see Figure 14c), although there are some scenarios in which the model's predictions are off (see, e.g., Figure 13 scenarios 4 and 7).

General Discussion

People are bombarded with social evaluations in their everyday lives. When Alice blames Bob for something bad that happened, a listener can readily infer that Alice is upset with Bob, and that Bob did something wrong. But, it turns out that social evaluations reveal much more than that. Decades of research have uncovered factors that drive responsibility judgments (Cushman, 2008; Gerstenberg et al., 2018; Lagnado et al., 2013; Langenhoff et al., 2021; Malle, 2021; Malle et al., 2014; Pizarro, Uhlmann, & Salovey, 2003; Shaver, 1985; Sosa et al., 2021; Uhlmann et al., 2015; Waldmann, Nagel, & Wiegmann, 2012; Weiner, 1995). Much of this research manipulates these factors to understand how they affect responsibility judgments. In this paper, we turned the tables around. We provided participants information about who was held responsible and how much, and asked them to infer what happened. Our work builds on prior research that looked at people's inferences about what happened based on physical changes to the environment (Chen & Scholl, 2016; Gerstenberg, Siegel, & Tenenbaum, 2021; Lopez-Brau et al., 2022; Pelz et al., 2020; Schachner & Kim, 2018), based on emotional expressions (Houlihan, Kleiman-Weiner, Hewitt, Tenenbaum, & Saxe, 2023; Ong et al., 2019; Saxe & Houlihan, 2017; Wu et al., 2021), and based on punishments and explanations (Kirfel et al., 2022; Radkani & Saxe, 2023). It is the first, to our knowledge, that investigates how people can draw rich inferences about what happened from *social evaluations*.

Across three experiments, we demonstrate that social evaluations provide a rich source of information. In Experiment 1, participants read a wide variety of vignettes in which one person was blamed more than another for an unfavorable outcome. Knowing who was blamed more allowed participants to make inferences about each person's ability, effort, knowledge, intention, and social role. Experiment 1 had broad coverage: it considered many different situations and asked participants to make inferences about several factors that had been shown to systematically affect responsibility judgments. However, the predictions based on these factors were merely qualitative: we correctly predicted which person participants selected in the scenario based on the blame information, but we weren't able to predict how strong their preference would be.

In Experiment 2, we developed and tested a quantitative model of praise focusing on one of the factors from Experiment 1: effort. In Experiment 2a, participants viewed a variety of maps that showed what path a person took to pick up a coffee for their co-worker. We asked participants how much praise the person deserved for picking up the coffee. We found that participants' praise judgments were well-predicted by a model that computes the additional effort that it took for the person to pick up the coffee compared to just having gone to the office directly. A person who goes out of their way to pick up coffee demonstrates that they care about you (as they're willing to incur a cost for your benefit) and are thus deserving of praise. Experiment 2b explored whether participants could infer what happened from knowing how much praise a person had received. We found that participants' inferences were well-accounted for by the praise models, and somewhat more consistent with having considered the actual effort that the person exerted rather than the additional effort compared to just going to the office directly.

In Experiment 3, we looked at a setting where several agents needed to coordinate their actions to bring about an optimal group outcome. In Experiment 3a, we asked

participants to judge what an agent should do. A model that assumes agents recursively think about each other's actions predicted participants' judgments well. In Experiment 3b, participants judged the extent to which each agent was to blame for a suboptimal group outcome. A model that considers what an agent should have done and how much their action affected the outcome, accurately captured participants' blame judgments. Finally, in Experiment 3c, participants had to infer what happened based on information about how much each agent was blamed. We found that participants' inferences were consistent with having considered both aspects of the blame model: whether agents had acted rationally and whether their action would have been pivotal for achieving the optimal group outcome.

Together, the results of these experiments demonstrate that people are capable of drawing sophisticated inferences from social evaluations, including both judgments of blame for negative outcomes and judgments of praise for positive ones. Experiment 1 provided qualitative support for such inferences across a wide variety of situations, and Experiments 2 and 3 complemented this with quantitative results, showing a close fit between generative models of praise and blame, and participants' judgments and inferences in two very different task environments. In the remainder, we want to highlight a few key observations, discuss limitations (see Table 4), point out fruitful avenues for future research, and draw out some broader implications of this work.

Unifying models of blame and praise

The two computational models we developed to capture responsibility judgments in Experiments 2 and 3 were quite different. Experiment 2 featured a relatively simple setting with a single agent who put in different amounts of effort to do a colleague a favor. Experiment 3 featured a group coordination setting where agents had to decide what action to take to achieve an optimal group outcome. Experiment 2 looked at praise and Experiment 3 at blame. Given the large differences between the setups in the two experiments, it's perhaps not surprising that different model implementations were needed. It is also not clear that the exact same principles hold for how people assign praise (Anderson et al., 2020) versus blame (Malle et al., 2014). That said, it would be nice to develop an even more unified account of how people generate social evaluations and make inferences based on them.

As we mentioned in the introduction, prior work argued that there are two key considerations that influence social evaluations: what causal role an agent's action played, and what the action reveals about the person (Gerstenberg, 2024; Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa, 1993; Wu et al., 2023). In the coffee scenarios, the agent's causal role is obvious. It's clear that the co-worker wouldn't have gotten a coffee without them. So the agent's action is always pivotal for generating the outcome. What this experiment manipulated is what one can learn about the agent from their action. How much effort the agent exerted reveals how much they value their co-worker. In the fishermen scenario, we saw that both components mattered. Agents were held more to blame when they took a suboptimal action, and when their action was (close to being) pivotal. Taking a suboptimal action could indicate that the agent didn't really care about the outcome, or that they weren't willing to put in the (cognitive) effort to think things through before acting. Because the outcome depended on the actions of several agents, this allowed for more nuanced differences in each agent's causal role. We found that agents received more

Table 4

Limitations of the present research.

Limitation	Explanation of limitation	Opportunity for future research
Model mismatch	The model that best accounts for participants praise judgments in Experiment 2a is not the model that best accounts for participants' inferences in Experiment 2b.	Better understanding when and how people might use simpler models for inference than for prediction.
Model disparity	The praise model for Experiment 2 is different than the blame model for Experiment 3.	To develop a unified model of blame and praise that applies across a wide variety of situations.
Explicit judgments	Participants made inferences based on explicit judgments of praise and blame.	Explore the subtle ways in which we express social evaluations in our everyday lives.
Single evaluations	Participants made inferences based on a single instance.	Explore people's inferences from social evaluations when more social evaluations are available (e.g., by multiple sources for the same instance, or by the same source for multiple instances).
Not all factors considered	The experiments only manipulate a subset of the factors that are known to influence social evaluations.	Investigate whether people can also draw inferences about other aspects such as what obliigations and norms hold in the situation, etc.
Additive combination assumption	Our computational models assume that causal attributions and person inferences combine additively to affect responsibility judgments.	It's plausible that these factors are related to one another in more intricate ways.
Inferences in the wild	We only studied inferences from social evaluations in controlled experiments.	Study how inferences from social evaluations arise in the wild.
Only adult participants	Our experiments only featured adult participants.	Explore how inferences from social evaluations develop with age.
Causal structure known	In all of our experiments, the causal structure of the situation was known.	Explore how people make inferences about the causal structure from social evaluations.
No inference about speaker	In our experiments, we provide little information about the speaker who produced the social evaluation.	Social evaluations are not just informative about what happened, they also license inferences about the speaker's mental model, including their beliefs and desires.

blame when their action was pivotal, such that the optimal group outcome could have been reached if they had acted differently. Overall, we believe that considering an agent's causal role and what the action reveals about them provides a unifying picture of how social evaluations are generated.

The results reported here extend prior work on responsibility judgments in groups. Lagnado et al. (2013) had shown that people's responsibility judgments were sensitive to

how close a person's contribution was to having been pivotal, and to how critical their contribution was perceived to have been before any actions took place. Gerstenberg et al. (2023) showed that people's judgments of how critical a person is for the outcome, are well captured by a model that assumes that people anticipate how much responsibility the person would bear for a positive outcome. For example, an individual's contribution in a conjunctive scenario – where each person needs to perform well in order for the group to succeed – is more critical than in a disjunctive scenario – where one person's good performance is sufficient for the group to succeed. So, overall, judgments of responsibility are sensitive both to a prospective component (criticality) and a retrospective component (pivotality; see also Engl, 2022). Here, our pivotality model that captures the retrospective component was similar to that of prior work (see Gerstenberg & Lagnado, 2010; Langenhoff et al., 2021; Wu & Gerstenberg, 2023; Zultan et al., 2012). However, the prospective component was different. Prior work looked at situations in which the individual contributions of the group members were largely independent from one another. In contrast, in our fishermen scenario, the different fishermen had to reason about what the other ones were likely to do. We saw that the recursive reasoning model captured participants' intuitions about what a person should do and that both components – what a person should have done and how much it mattered – jointly affected responsibility judgments.

While the two models that generated praise and blame predictions in Experiment 2 and 3 differed, the way in which inference was carried out was largely identical. In both cases, we assumed that people map from continuous judgments to discrete labels to determine whether a person was deserving of "low", "medium", or "high" praise or blame. Then, we assumed that people consider the different possible ways in which the situation could have unfolded and what praise (or blame) judgments would have been most appropriate in each situation. Participants' inferences were consistent with assuming that they performed Bayesian inference over the different possibilities by conditioning on the evidence they had received.

Normative versus descriptive judgments and inferences

Normative models that dictate how behavior should be explained (e.g., Kelley, 1973; Morris & Larrick, 1995) and responsibility attributed (e.g., Shaver, 1985) have a long tradition in psychology (see Alicke et al., 2015, for a review). For example, Kelley's (1973) ANOVA model prescribes what factors people should ascribe behavior to based on covariation information about the actor, other actors, and situations. Shaver's (1985) theory of blame prescribes that people should first establish causation, then assess responsibility by considering the actor's mental states, and finally assign blame based on whether the actor had justifications or excuses for their behavior. However, human judgments sometimes fall short of these normative standards. People may underestimate the importance of situational pressures and overemphasize personal factors when attributing behavior (Gilbert & Malone, 1995; Jones & Harris, 1967; Ross, 1977 although see Walker, Smith, & Vul, 2022). They may also exaggerate an actor's causal role to validate the blame they would like to give (Alicke, 2000). Our work here builds on this prior tradition. In Experiment 1, we showed that factors that have been argued to have normative import for how people should ascribe responsibility can be inferred from judgments of blame (Shaver, 1985). And in Experiments 2 and 3, we showed that participants' praise and blame judgments, as well

as their inferences from these judgments, were consistent with models that consider what causal role and agent's action played, and what the action reveals about the person (see also Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa et al., 2021; Wu et al., 2023).

There are two ways in which questions about normativity arise in our studies. First, whether people's responsibility judgments are (normatively) accurate and, second, whether their inferences based on responsibility judgments are accurate. We don't offer a normative account of people's responsibility judgments here. That account would have to explain why it's good that people judge responsibility the way they do, for example, by postulating what consequences judgments of responsibility (and the anticipation of these judgments) have for individuals and society. As mentioned earlier, one such justification could come from the effects that moral and social evaluations have on how we regulate relationships (Rai & Fiske, 2011). For predicting how people make inferences from others' responsibility judgments, we relied on Bayes' theorem: a normative tool for modeling belief change in light of evidence. It certainly is possible within this framework, for mismatches between responsibility judgments and inferences to arise. For example, someone might give more praise than a person deserves in order to promote positive future behavior, and a person who learns about that praise (but doesn't know about the further motivations of the praise-giver) makes an incorrect inference about what happened.[14] Bayesian inference relies on evaluating the evidence in light of all hypotheses that could have produced the data, but it's plausible that people only ever consider a small subset of hypotheses (Bramley, Zhao, Quillien, & Lucas, 2023; Lieder & Griffiths, 2020). In fact, we observed such a mismatch between generative and inferential model in in Experiment 2. Here, the praise judgments of most participants in the coffee paradigm were best explained by a model that considered the additional cost that a person incurred by picking up a coffee for their co-worker. However, when asked to infer what happened based on how much praise the person had received, participants' inferences were more in line with having considered the action cost itself rather than the additional action cost.

One possible explanation for the mismatch between the judgment and inference task is that inferences are computationally more demanding. In the coffee experiment, the actual effort that each agent incurred can be easily read off the maps. In contrast, how much additional effort each agent put in (compared with going directly to the office without picking up the coffee) requires computation. While this calculation is also required when making praise judgments, it's easier in this setting because only one situation needs to be considered. The inference setting demands this computation for two maps before judging which situation is more likely given the praise judgment. More work is needed to better understand how resource limitations influence how participants make social evaluations and learn from them (Alanqary et al., 2021; Icard, 2023; Levine, Chater, Tenenbaum, & Cushman, 2023; Lieder & Griffiths, 2020). That said, it's worth noting that the differences between the model fits were relatively small and it's of course possible that factors other than computational complexity may have contributed to differences in the models that best captured participants' judgments versus their inferences.

Relatedly, it would be interesting to explore how much of this knowledge is explicit versus implicit. Recall that in Experiment 3a, we found that some participants' judgments

[14] We thank anonymous reviewer for making this point.

about what the fisherman should do were best predicted by a recursive reasoning model, whereas for others, the heuristic model captured their judgments best. It's possible that participants differed in how carefully they thought through the different situations and this difference in mental computation was reflected in their judgments. More generally, do people know what factors influence their blame and praise judgments? In principle, explicit knowledge is not required to infer what happened. People just need to know *what* kinds of social evaluations they would have given in different situations, without needing to know *why* they would have done so.

Making inferences about the speaker from responsibility judgments

In our experiments, we provide little information about the speaker who generated the responsibility judgment (see Figure 1). However, knowledge about the speaker affects what inferences one can draw (Schuster & Degen, 2020; Yildirim, Degen, Tanenhaus, & Jaeger, 2016). Even children quickly adjust their inferences by learning speaker-specific patterns of speech, such as how judicious they are with their praise (Asaba & Gweon, 2020; Asaba et al., 2018; Lee, Kim, Kesebir, & Han, 2017; Yoon et al., 2020). For example, if Tom received a lot of praise for his performance from Betty, it matters whether Betty is Tom's mother or an independent observer. We know that parents tend to overpraise their children, so a raving review from a mother leaves some doubts about how good the performance actually was. So, whenever a person expresses a judgment of responsibility, that judgment can reveal both something about what happened but also information about the speaker. In cases where the listener already knows what happened, a speaker's responsibility judgment may reveal why the speaker thought it happened and what they think should be done about it (Kirfel et al., 2024; Sehl, Denison, & Friedman, 2024). These inferences happen not only in relatively mundane cases of parents overpraising their children, but also in much more severe cases that have large societal implications. For instance, the US has a school shooting epidemic, with sharp disagreements on who and what is to blame. While some people (or news media) may blame what happened on the shooter's mental health, others will blame it on the availability of guns. Depending on what cause a speaker cites, the listener can infer the speaker's desire for what should or shouldn't be changed in society.

In addition to their motivations, what a speaker chooses to say can also reveal their beliefs. In our experiments, we assumed that people knew how things worked, but didn't know what happened. A person's judgment can reveal their mental model of the situation. For example, if someone blames the wrong person for a team failure, this suggests that they may have an incorrect causal model of the situation. They may have thought that this person's contributions were critical when, in fact, they didn't matter for the outcome (Gerstenberg et al., 2023). In this way, social evaluations give us information not only about what happened but also about the speaker's mental model of the situation, including their beliefs and desires. A challenge for social cognition research is to better understand how people make these joint inferences – the framework we developed here provides a good starting point.

Limitations & Future Directions

Social evaluations take many different forms. We focused here on judgments of blame and praise. In our experiments, these judgments were explicitly communicated. In Experiment 1, we stated who had received more (or less) blame, and in Experiments 2b and 3c, we stated that an agent had received "low", "medium", or "high" blame. In our everyday lives, social evaluations are often communicated less directly. Additional work would be required to translate people's everyday expressions into a format such that the models we considered here would work. One possibility would be a hybrid approach that combines the flexibility of large language models for handling various kinds of inputs with the principled ways of making inferences that Bayesian generative model afford (see, e.g., Wong et al., 2023). While our participants were capable of drawing inferences from social evaluations in experimental settings, one might wonder how often these kinds of inferences arise in the wild. We believe that they may be quite prominent. For example, news headlines often include social evaluations that may shape a reader's beliefs about what happened. When friends share gossip, they may sometimes begin with a social evaluation that creates an early impression on the listener. Future work should study spontaneous inferences from social evaluations (e.g., Schneid, Carlston, & Skowronski, 2015), as well as how the ability to make such inferences develops with age (Amemiya et al., 2024; Jara-Ettinger et al., 2015).

We considered only a subset of the factors that have been shown to influence social evaluations. We didn't look at the role of norms (Hamilton, 1978; Hamilton, Blumenfeld, & Kushler, 1988; Hitchcock & Knobe, 2009; Malle, 2021), for example, or at how the causal structure of the situation affects responsibility judgments in groups (Gerstenberg & Lagnado, 2010; Gerstenberg et al., 2023; Lagnado et al., 2013; Zultan et al., 2012). Moreover, we assumed that the factors that we identified combine additively to affect responsibility judgments. For example, the computational model in Experiment 3 assumes that people assign blame by considering how rational a person's action was and how pivotal it was (see Equation 10). However, it's plausible that these factors interact with one another. In order to be held responsible, one's action has to have had some causal connection with the outcome (Shaver, 1985, see). An irrational action that had nothing to do with an outcome, cannot be blamed for it. Our experiments didn't include agents whose actions didn't affect the outcome, but this could be explored in future work. Petrocelli, Percy, Sherman, and Tormala (2011) developed a model of counterfactual potency according to which counterfactual thoughts are potent when it's easy to imagine that things could have been different (what they term the "if-likelihood"), and when they would have led to a different outcome from what actually happened (the "then-likelihood"). In their model, counterfactual potency is a multiplicative combination of if-likelihood and then-likelihood. More work is needed to explore exactly how different factors combine to affect responsibility attributions and inferences. Finally, in our experiments, participants made inferences from a single social evaluation. Future work should explore what inferences people can make from multiple, potentially conflicting social evaluations (Amemiya et al., 2024, 2021).

Implications of this work

A computational understanding of how people draw inferences from social evaluations has implications for several areas of study. First, it will inform work in computer

science on artificial intelligence and human-computer interaction. Having accurate models of how people make social judgments and what they can learn from them is critical for building and evaluating machines that emulate such human capabilities. Second, understanding what inferences people draw from social evaluations is relevant for research into how people make decisions in social environments. When deciding what to do, people are likely to be motivated by the judgments and inferences that others are likely to make. By better understanding inferences from social evaluations, we'll be able to build more accurate models of how people will act in social situations. Finally, this work may also have implications for legal studies. Witness statements often include social evaluations, and it's important to better understand what inferences jury members are likely to draw from such statements (Lagnado & Gerstenberg, 2017; Summers, 2018).

Conclusion

Social evaluations, such as judgments of blame or praise, are part of the fabric of people's lives. Past work has helped us learn more about how people make such judgments. This paper turns things around and asks what people can infer from others' social evaluations. 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 have only just begun to better understand scientifically.

Acknowledgments

We thank the members of the Causality in Cognition Lab (CiCL) for feedback and discussion. TG was supported by grants from Stanford's Human-Centered Artificial Intelligence Institute (HAI) and from Cooperative AI. JJE was supported by NSF award BCS-2045778. Some of the results from Experiment 3 have been reported in Allen et al. (2015) (Experiment 3a and 3b) and in Davis et al. (2021) (Experiment 3c).

References

Alanqary, A., Lin, G. Z., Le, J., Zhi-Xuan, T., Mansinghka, V. K., & Tenenbaum, J. B. (2021). Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv preprint arXiv:2106.13249*.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.

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 D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 84–89). Austin, TX: Cognitive Science Society.

Amemiya, J., Heyman, G. D., & Gerstenberg, T. (2024). Children use disagreement to infer what happened. *Cognition*.

Amemiya, J., Walker, C. M., & Heyman, G. D. (2021). Children's developing ability to resolve disagreements by integrating perspectives. *Child Development*, *92*(6), e1228–e1241.

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in cognitive sciences*, *24*(9), 694–703.

Asaba, M., & Gweon, H. (2020). Learning about others to learn about the self: Early reasoning about the informativeness of others' praise. In *Psychological perspectives on praise* (pp. 67–74). Routledge.

Asaba, M., Hembacher, E., Qiu, H., Anderson, B., Frank, M. C., & Gweon, H. (2018). Young children use statistical evidence to infer the informativeness of praise. In *Cogsci.*

Baker, C., Tenenbaum, J., & Saxe, R. (2006). Bayesian models of human action understanding. *Advances in Neural Information Processing Systems*, *18*, 99-106.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 749–753.

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

Beller, A., Xu, Y., Linderman, S., & Gerstenberg, T. (2022). Looking into the past: Eye-tracking mental simulation in physical inference. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society* (pp. 3641–3647). Cognitive Science Society.

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*(12), 1654.

Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (2023). Local search and the

evolution of world models. *Topics in Cognitive Science*.

Brewer, M. B. (1977). An information-processing approach to attribution of responsibility. *Journal of Experimental Social Psychology*, *13*(1), 58–69.

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

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

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.

Costa-Gomes, M., Crawford, V. P., & Broseta, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica*, *69*(5), 1193–1235.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS One*, *4*(8), e6699.

Davis, Z. J., Allen, K. R., & Gerstenberg, T. (2021). Who went fishing? inferences from social evaluations. In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society.*

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dunlea, J. P., & Heiphetz, L. (2020). Children's and adults' understanding of punishment and the criminal justice system. *Journal of Experimental Social Psychology*, *87*, 103913.

Engl, F. (2022). Causal responsibility attribution: Theory and experimental evidence.

Felsenthal, D. S., & Machover, M. (2009). A note on measuring voters' responsibility. *Homo Oeconomicus*, *26*(2), 259–271.

Fincham, F. D., & Jaspars, J. M. (1983). A subjective probability approach to responsibility attribution. *British Journal of Social Psychology*, *22*(2), 145–161.

Gerstenberg, T. (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*. (https://osf.io/preprints/psyarxiv/72scr)

Gerstenberg, T., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. In C. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 720–725). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(6), 936–975.

Gerstenberg, T., & Icard, T. F. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of respon-

sibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review*, *19*(4), 729–736.

Gerstenberg, T., Lagnado, D. A., & Kareev, Y. (2010). The dice are cast: The role of intended versus actual contributions in responsibility attribution. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1697–1702). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Lagnado, D. A., et al. (2023). Making a positive difference: Criticality in groups. *Cognition*, *238*, 105499.

Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. B. (2021). What happened? reconstructing the past from vision and sound. *PsyArXiv*.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

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.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38.

Gilbert, E. A., Tenney, E. R., Holland, C. R., & Spellman, B. A. (2015). Counterfactuals, control, and causation: Why knowledgeable people get blamed more. *Personality and Social Psychology Bulletin*, *41*(5), 643–658.

Guglielmo, S., & Malle, B. F. (2010). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition*, *117*(2), 139–150.

Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the thirty-second aaai conference on artificial intelligence (aaai-18)* (pp. 1853–1860).

Hamilton, V. L. (1978). Who is responsible? Toward a social psychology of responsibility attribution. *Social Psychology*, *41*(4), 316–328.

Hamilton, V. L., Blumenfeld, P. C., & Kushler, R. H. (1988). A question of standards: Attributions of blame and credit for classroom acts. *Journal of Personality and Social Psychology*, *54*(1), 34.

Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. New York: Oxford University Press.

Hertwig, R., & Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, *11*(3), 359–372.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, *11*, 587–612.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*.

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.

Icard, T. F. (2023). Resource rationality.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

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*(10), 785.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Johnson, S. G., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, *77*, 42–76.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153.

Keith, M. G., & McKay, A. S. (2024). Too anecdotal to be true? mechanical turk is not all bots and bad data: Response to webb and tangney (2022). *Perspectives on Psychological Science*, 17456916241234328.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128.

Kirfel, L., Bunk, X., Zultan, R., & Gerstenberg, T. (2023). Father, don't forgive them, for they could have known what they're doing. In M. B. Goldwater, F. Anggoro, B. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Kirfel, L., Harding, J., Shin, J., Xin, C., Icard, T., & Gerstenberg, T. (2024). Do as I explain: Explanations communicate optimal interventions. In L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481–1501.

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Kirfel, L., & Lagnado, D. A. (2019). I know what you did last summer (and how often). epistemic states and statistical normality in causal judgments. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Kirfel, L., & Phillips, J. (2023). The pervasive impact of ignorance. *Cognition*, *231*, 105316.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1123–1128). Austin, TX: Cognitive Science Society.

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

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 676–681). Austin, TX: Cognitive Science Society.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315–365.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Lagnado, D. A., & Gerstenberg, T. (2015). A difference-making framework for intuitive judgments of responsibility. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (Vol. 3, pp. 213–241). Oxford University Press.

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.

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

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, 101412.

Lee, H. I., Kim, Y.-H., Kesebir, P., & Han, D. E. (2017). Understanding when parental praise leads to optimal child outcomes: Role of perceived praise accuracy. *Social Psychological and Personality Science*, *8*(6), 679–688.

Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2023). Resource-rational contractualism: A triple theory of moral cognition.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

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

Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley.

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

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science*, *31*(3), 227–242.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331–355.

Naumov, P., & Tao, J. (2018). Blameworthiness in games with imperfect information. *CoRR*.

Navarre, N., Konuk, C., Bramley, N. R., & Mascarenhas, S. (2024). Functional rule inference from causal selection explanations. In L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the*

Cognitive Science Society.

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.

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).

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology*, *100*(1), 30–46.

Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychological Science*, *14*(3), 267–72.

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

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Radkani, S., & Saxe, R. (2023). What people learn from punishment: joint inference of wrongness and punisher's motivations from observation of punitive choices. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Radkani, S., Tenenbaum, J., & Saxe, R. (2022). Modeling punishment as a rational communicative social action. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75.

Richens, J., Beard, R., & Thompson, D. H. (2022). Counterfactual harm. *Advances in Neural Information Processing Systems*, *35*, 36350–36365.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173–220.

Sarin, A., & Cushman, F. (2024). One thought too few: An adaptive rationale for punishing negligence. *Psychological Review*, *131*(3), 812.

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544.

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a bayesian model of theory of mind. *Current Opinion in Psychology*, *17*, 15 - 21. (Emotion)

Schachner, A., & Kim, M. (2018). Alternative causal explanations for order break the link between order and agents. *PsyArXiv*.

Schelling, T. C. (1980). *The strategy of conflict: with a new preface by the author*. Harvard university press.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review*, *101*(4), 632–652.

Schneid, E. D., Carlston, D. E., & Skowronski, J. J. (2015). Spontaneous evaluative inferences and their relationship to spontaneous trait inferences. *Journal of Personality*

and Social Psychology, *108*(5), 681.

Scholten, M. (2022). Blaming friends. *Philosophical Studies*, *179*(5), 1545–1562.

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, 104285.

Sehl, C. G., Denison, S., & Friedman, O. (2024). Doing things efficiently: Testing an account of why simple explanations are satisfying. *Cognitive Psychology*, *154*, 101692.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer-Verlag, New York.

Shu, T., Kryven, M., Ullman, T. D., & Tenenbaum, J. B. (2020). Adventures in flatland: Perceiving social interactions under physical dynamics. In *42d proceedings of the annual meeting of the cognitive science society*.

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

Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1467–1472). Austin, TX: Cognitive Science Society.

Sosa, D. (1993). Consequences of consequentialism. *Mind*, *102*(405), 101–122.

Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, *217*, 104890.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323–348.

Stahl II, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of economic behavior & organization*, *25*(3), 309–327.

Summers, A. (2018). Common-sense causation in the law. *Oxford Journal of Legal Studies*, *38*(4), 793–821.

Sytsma, J. (2021). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*, *12*, 699–719.

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

Ullman, T. D., Tenenbaum, J. B., Baker, C. L., Macindoe, O., Evans, O. R., & Goodman, N. D. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems* (Vol. 22, pp. 1874–1882).

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In *The oxford handbook of thinking and reasoning* (pp. 364–389). New York: Oxford University Press.

Walker, D., Smith, K. A., & Vul, E. (2022). Reconsidering the "bias" in "the correspondence bias". *Decision*, *9*(3).

Webb, M. A., & Tangney, J. P. (2022). Too good to be true: Bots and bad data from mechanical turk. *Perspectives on Psychological Science*, 17456916221120027.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psy-*

chological Review, *92*(4), 548.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* New York: The Guilford Press.

Weiner, B., & Kukla, A. (1970). An attributional analysis of achievement motivation. *Journal of Personality and Social Psychology*, *15*(1), 1–20.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv*.

Wu, S. A., Brockbank, E., Cha, H., Fränken, J.-P., Jin, E., Huang, Z., ... Gerstenberg, T. (2024). Whodunnit? Inferring what happened from multimodal evidence. In L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Wu, S. A., & Gerstenberg, T. (2023). If not me, then who? Responsibility and replacement. *PsyArXiv*.

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences. In M. B. Goldwater, F. Anggoro, B. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Wu, Y., Muentener, P., & Schulz, L. E. (2017). One- to four-year-olds connect diverse positive emotional vocalizations to their probable causes. *Proceedings of the National Academy of Sciences*, 201707715.

Wu, Y., Schulz, L. E., Frank, M. C., & Gweon, H. (2021). Emotion as information in early social learning. *Current Directions in Psychological Science*, *30*(6), 468–475.

Xiang, Y., Landy, J., Cushman, F., Vélez, N., & Gershman, S. J. (2023). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71–87.

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.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition*, *125*(3), 429–440.

Appendix A
Experiment 1

Scenarios

In the experiment, we randomly sampled the names for *Requester*, *Person 1* and *Person 2* from the top 25 male and female names of the past 100 years and adapted the pronouns in the stories accordingly.

Ability

Scenario 1:

Requester needed help fixing their flat tire. *Requester* asked *Person 1* and *Person 2* for help, but neither did. *Requester* didn't manage to fix the tire by themselves.
Requester blamed *Person 1* more/less than *Person 2* for not helping.

Who is better/worse at fixing tires?

Scenario 2:

Requester needed help carrying a very heavy couch. *Requester* asked *Person 1* and *Person 2* for help, but both said no. *Requester* wasn't able to move the couch on their own.
Requester blamed *Person 1* more than *Person 2* for not helping.

Who is stronger/weaker?

Scenario 3:

Requester needed help reaching something on a high shelf. *Requester* asked *Person 1* and *Person 2* for help, but both said no. *Requester* wasn't able to reach the item.
Requester blamed *Person 1* more/less than *Person 2* for not helping.

Who is taller/shorter?

Effort

Scenario 4:

Requester was on a tug of war team with *Person 1* and *Person 2*. Their team lost the game of tug of war.
Requester blamed *Person 1* more/less than *Person 2* for their loss.

Who tried / didn't try very hard?

Scenario 5:

Requester was assigned to do a group project with *Person 1* and *Person 2*. Their group did a bad job and received a failing grade for the assignment.
Requester blamed *Person 1* more/less than *Person 2* for the group's failing grade.

Who tried / didn't try very hard on the assignment?

Scenario 6:

Requester needed someone to make an appointment for them. *Requester* asked *Person 1* and *Person 2* to call and wait on hold for them. *Person 1* and *Person 2* both gave up before their call was picked up. *Requester* later learned that one of them waited for a few minutes, and another waited for hours.
Requester blamed *Person 1* more/less than *Person 2* for hanging up.

Who waited for hours / a few minutes before giving up?

Knowledge

Scenario 7:

Requester texted *Person 1* and *Person 2* to pick them up from the airport. Neither responded to their text. *Requester* later learned that only one of them saw their message.
Requester blamed *Person 1* more/less than *Person 2* for not helping.

Who saw / didn't see the text message?

Scenario 8:

Person 1, *Person 2*, and *Requester* were having a picnic. Both *Person 1* and *Person 2* brought snacks with peanuts in them, even though *Requester* was allergic to peanuts.
Requester blamed *Person 1* more/less than *Person 2* for bringing snacks with peanuts to the picnic.

Who knew / didn't know that *Requester* was allergic to peanuts?

Scenario 9:

Requester needed help making dinner. Neither *Person 1* nor *Person 2* helped.
Requester blamed *Person 1* more than *Person 2* for not helping.

Who knew / didn't know that *Requester* needed help making dinner?

Intention

Scenario 10:

Requester fell over while playing soccer with *Person 1* and *Person 2*. Both *Person 1* and *Person 2* stepped on *Requester*.
Requester blamed *Person 1* more/less than *Person 2* for stepping on them.

Who stepped on *Requester* on purpose / by accident?

Scenario 11:

Requester always packs themselves a lunch to bring to work. One day, *Person 1* ate *Requester*'s lunch. A different day, *Person 2* ate *Requester*'s lunch.
Requester blamed *Person 1* more/less than *Person 2* for eating their lunch.

Who ate *Requester*'s lunch on purpose / accidentally?

Scenario 12:

Person 1 and *Person 2* each had a birthday party in separate weeks. Neither invited *Requester* to the party. *Requester* later learned from a friend that one of them intentionally didn't invite them, and another forgot to invite them.
Requester blamed *Person 1* more/less than *Person 2* for not inviting them to Person 1's birthday party.

Who intentionally didn't invite / forgot to invite *Requester*?

Social role

Scenario 13:

Requester had a birthday party. *Requester* asked *Person 1* and *Person 2* to come, but neither came.
Requester blamed *Person 1* more than *Person 2* for not coming.

Who is *Requester*'s closer / less close friend?

Scenario 14:

Requester needed help moving. *Requester* asked *Person 1* and *Person 2* for help, but both said no. *Requester* wasn't able to move their stuff on their own. *Requester* blamed *Person 1* more than *Person 2* for not helping.

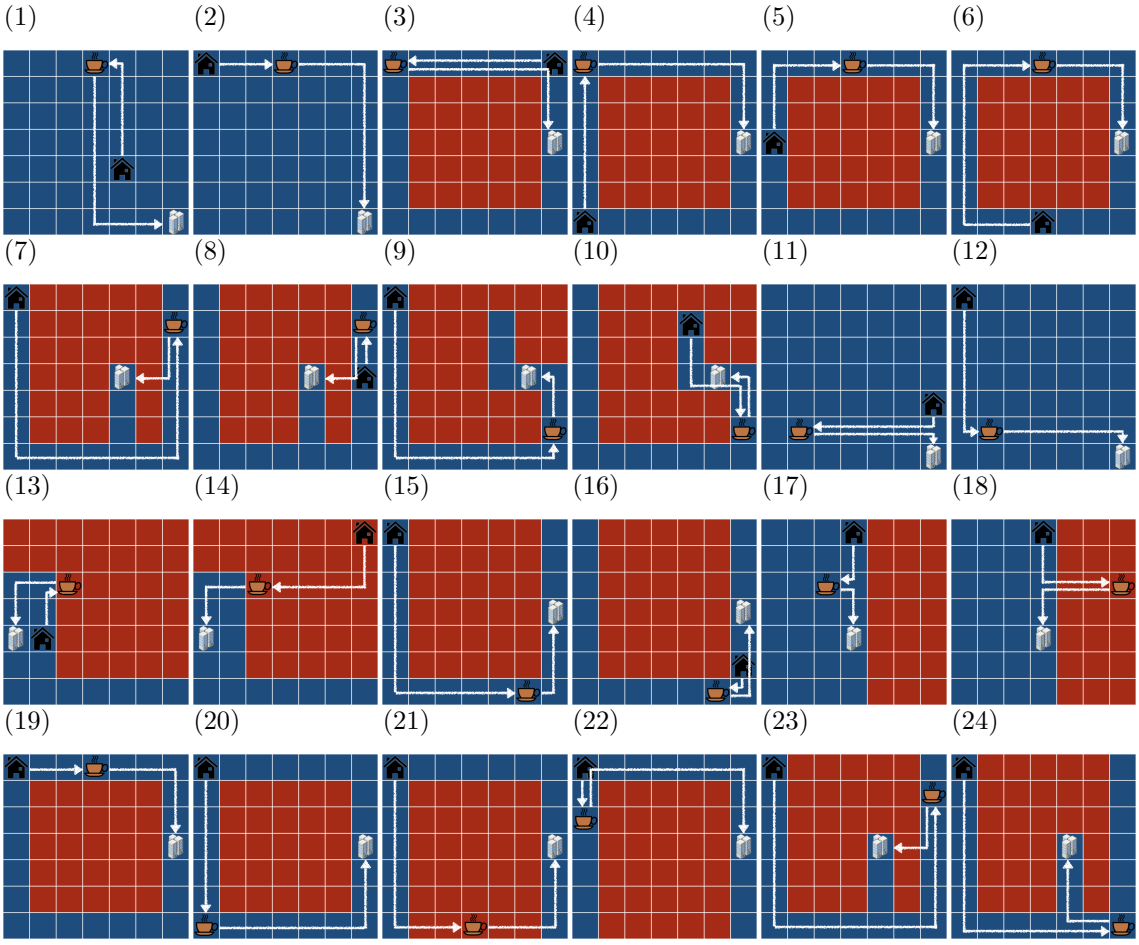
Who is *Requester*'s closer / less close friend?

Scenario 15:

Requester needed help practicing for the school play. *Requester* asked *Person 1* and *Person 2* for help, but both said no.
Requester blamed *Person 1* more than *Person 2* for not helping.

Who is *Requester*'s closer / less close friend?

Appendix B
Experiment 2



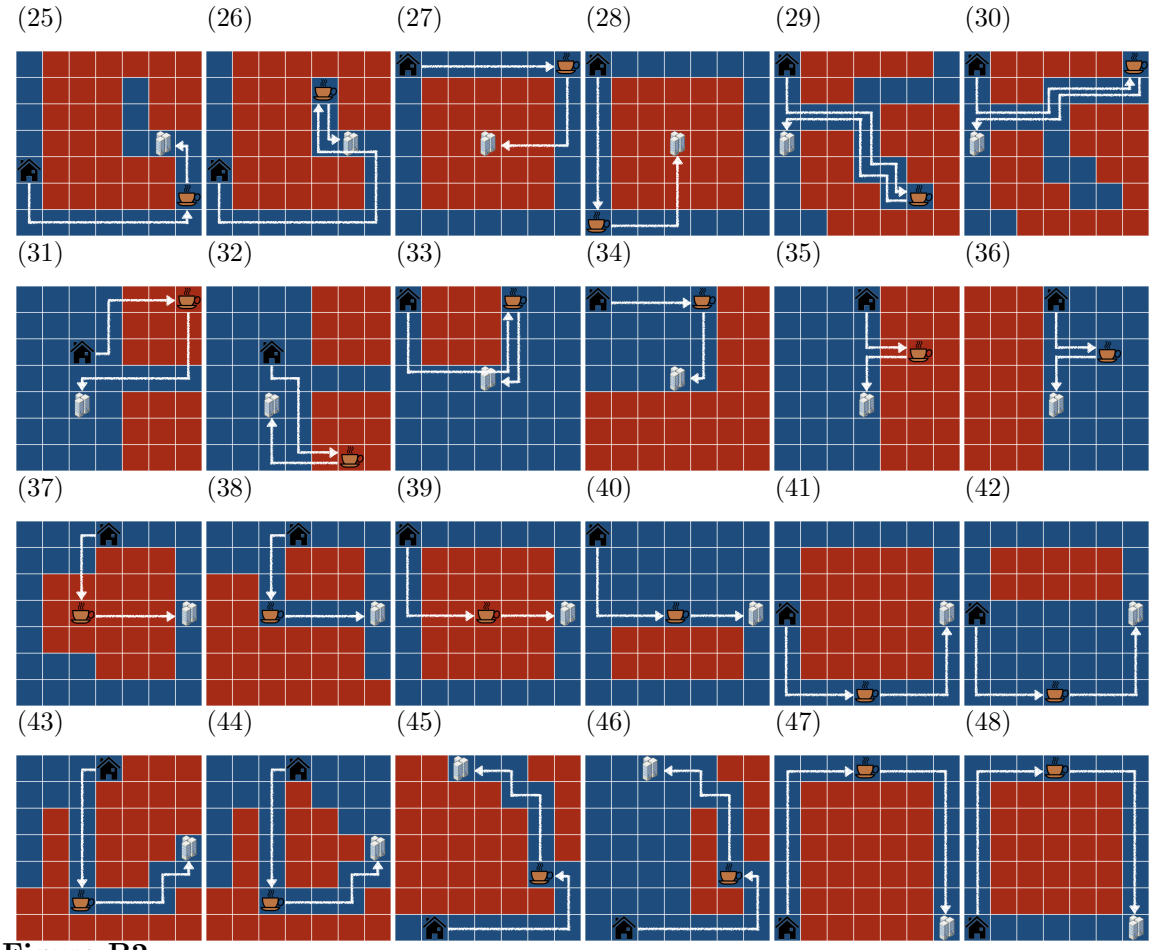


Figure B2

Experiment 2a trials.

Table B1

Experiment 2a*. Trial information with model predictions and participant judgments (means with 95% bootstrapped confidence intervals).* Note: *The map numbers refer to the ones in Figure B2.*

map	action cost	additional action cost	action cost model	additional action cost model	full model	judgments
1	14	10	46	56	55	59 (52, 65)
2	12	0	42	32	33	28 (21, 35)
3	15	12	48	61	60	64 (56, 71)
4	15	6	48	47	47	42 (35, 49)
5	12	0	42	32	33	32 (25, 38)
6	18	12	55	61	61	53 (46, 60)
7	23	10	65	56	60	61 (56, 67)
8	8	4	34	42	39	43 (38, 49)
9	16	0	51	32	35	31 (26, 38)
10	9	6	36	47	43	49 (43, 55)
11	12	10	42	56	54	62 (56, 68)
12	12	0	42	32	33	27 (20, 33)
13	9	8	36	51	48	59 (53, 65)
14	22	0	63	32	39	47 (38, 55)
15	15	0	48	32	35	30 (24, 36)
16	6	4	30	42	37	29 (23, 35)
17	6	2	30	37	33	30 (24, 36)
18	20	16	59	71	71	72 (67, 76)
19	9	0	36	32	31	26 (19, 33)
20	15	6	48	47	47	45 (38, 51)
21	25	16	69	71	74	75 (69, 80)
22	13	4	44	42	42	37 (31, 44)
23	23	10	65	56	60	60 (54, 66)
24	17	4	53	42	44	46 (40, 52)
25	12	0	42	32	33	33 (26, 40)
26	16	4	51	42	43	50 (44, 55)
27	18	6	55	47	49	56 (50, 61)
28	18	6	55	47	49	52 (46, 57)
29	19	16	57	71	70	64 (57, 71)
30	19	16	57	71	70	69 (62, 75)
31	24	22	67	85	86	75 (68, 81)
32	18	16	55	71	70	67 (61, 73)
33	14	8	46	51	51	54 (46, 61)
34	8	2	34	37	34	36 (30, 42)
35	14	10	46	56	55	62 (58, 67)
36	8	4	34	42	39	38 (32, 43)
37	18	12	55	61	61	70 (64, 75)
38	8	2	34	37	34	34 (28, 41)
39	19	10	57	56	58	68 (62, 74)
40	9	0	36	32	31	27 (22, 34)
41	12	0	42	32	33	29 (23, 36)
42	12	6	42	47	45	45 (38, 51)
43	12	0	42	32	33	34 (28, 40)
44	12	6	42	47	45	47 (41, 53)
45	15	0	48	32	35	33 (27, 40)
46	15	8	48	51	51	47 (41, 53)
47	18	2	55	37	41	40 (33, 47)
48	18	12	55	61	61	51 (44, 58)

Table B2

Experiment 2b. *Trial information with model predictions and participant judgments (means with 95% bootstrapped confidence intervals).* Note: *The map numbers refer to the ones in Figure B2.*

trial	map 1	map 2	praise	action cost model	additional action cost model	full model	empirical model	judgment
1	15	16	low	0.67	0.47	0.49	0.50	0.62 (0.57, 0.68)
2	23	12	low	0.90	0.66	0.71	0.82	0.87 (0.83, 0.90)
3	45	2	low	0.60	0.50	0.50	0.52	0.69 (0.65, 0.73)
4	47	25	low	0.72	0.51	0.52	0.55	0.75 (0.71, 0.79)
5	33	34	low	0.63	0.60	0.60	0.71	0.82 (0.78, 0.85)
6	45	46	low	0.50	0.40	0.40	0.37	0.46 (0.42, 0.51)
7	5	6	low	0.28	0.28	0.28	0.28	0.14 (0.11, 0.17)
8	19	20	low	0.35	0.44	0.43	0.39	0.18 (0.14, 0.22)
9	35	36	low	0.63	0.63	0.63	0.82	0.85 (0.81, 0.88)
10	37	38	low	0.77	0.71	0.72	0.92	0.85 (0.82, 0.89)
11	39	40	low	0.80	0.66	0.68	0.91	0.82 (0.78, 0.86)
12	21	22	low	0.94	0.83	0.86	0.95	0.85 (0.82, 0.89)
13	9	10	medium	0.41	0.65	0.58	0.62	0.56 (0.50, 0.60)
14	19	20	medium	0.59	0.65	0.66	0.68	0.63 (0.58, 0.67)
15	41	42	medium	0.50	0.65	0.62	0.64	0.55 (0.50, 0.60)
16	5	6	medium	0.51	0.69	0.66	0.61	0.57 (0.52, 0.62)
17	27	28	medium	0.50	0.50	0.50	0.51	0.54 (0.52, 0.57)
18	47	48	medium	0.50	0.63	0.58	0.53	0.54 (0.50, 0.58)
19	24	26	medium	0.50	0.50	0.50	0.51	0.53 (0.49, 0.58)
20	42	46	medium	0.53	0.53	0.53	0.51	0.62 (0.57, 0.66)
21	13	14	medium	0.50	0.33	0.43	0.52	0.37 (0.32, 0.42)
22	21	20	medium	0.69	0.50	0.53	0.67	0.60 (0.55, 0.65)
23	7	8	medium	0.50	0.42	0.40	0.52	0.45 (0.40, 0.50)
24	39	28	medium	0.51	0.46	0.48	0.60	0.65 (0.60, 0.69)
25	11	12	high	0.50	0.10	0.13	0.06	0.22 (0.18, 0.27)
26	17	18	high	0.95	0.91	0.93	0.93	0.89 (0.86, 0.92)
27	21	22	high	0.22	0.13	0.13	0.13	0.18 (0.14, 0.22)
28	37	38	high	0.11	0.12	0.11	0.11	0.16 (0.12, 0.20)
29	3	4	high	0.50	0.27	0.30	0.22	0.23 (0.19, 0.27)
30	23	24	high	0.37	0.24	0.24	0.31	0.19 (0.15, 0.22)
31	39	28	high	0.47	0.33	0.35	0.35	0.34 (0.29, 0.39)
32	13	14	high	0.90	0.13	0.31	0.33	0.60 (0.54, 0.65)
33	29	30	high	0.50	0.50	0.50	0.53	0.37 (0.33, 0.40)
34	31	32	high	0.40	0.44	0.43	0.48	0.20 (0.16, 0.25)
35	11	13	high	0.32	0.42	0.39	0.47	0.50 (0.45, 0.56)
36	35	1	high	0.50	0.50	0.50	0.47	0.51 (0.46, 0.56)

Appendix C

Experiment 3

Table C1

Experiment 3a. *Trial information with model predictions and participant judgments (means with 95% bootstrapped confidence intervals).*

index	trees	A	B	C	recursive reasoning	heuristic choice	judgment
1	1	1	2	2	0.95	0.95	0.94 (0.87, 0.98)
2	1	2	2	2	0.35	0.43	0.42 (0.29, 0.57)
3	1	3	1	2	0.00	0.17	0.02 (0.01, 0.03)
4	1	2	3	3	1.00	0.95	0.87 (0.77, 0.95)
5	1	2	1	1	0.08	0.17	0.08 (0.02, 0.17)
6	1	1	3	3	1.00	0.95	0.91 (0.81, 0.98)
7	1	3	3	3	0.26	0.43	0.36 (0.25, 0.5)
8	1	1	1	2	0.67	0.56	0.78 (0.66, 0.89)
9	1	2	1	3	0.04	0.17	0.12 (0.04, 0.23)
10	2	1	2	2	0.15	0.17	0.2 (0.08, 0.35)
11	2	2	2	2	0.35	0.43	0.56 (0.42, 0.7)
12	2	3	1	2	0.05	0.17	0.06 (0.02, 0.13)
13	2	2	3	3	1.00	0.95	0.84 (0.71, 0.95)
14	2	2	1	1	0.69	0.56	0.79 (0.66, 0.91)
15	2	1	3	3	0.09	0.17	0.16 (0.06, 0.28)
16	2	3	3	3	0.26	0.43	0.4 (0.25, 0.53)
17	2	1	1	2	0.42	0.56	0.34 (0.19, 0.5)
18	2	2	1	3	0.97	0.95	0.89 (0.8, 0.96)
19	3	1	2	2	0.71	0.95	0.78 (0.64, 0.91)
20	3	2	2	2	0.61	0.69	0.59 (0.46, 0.7)
21	3	3	1	2	0.86	0.56	0.82 (0.69, 0.93)
22	3	2	3	3	0.01	0.17	0.15 (0.06, 0.28)
23	3	2	1	1	0.61	0.95	0.9 (0.8, 0.97)
24	3	1	3	3	0.09	0.17	0.13 (0.04, 0.25)
25	3	3	3	3	0.26	0.43	0.56 (0.4, 0.72)
26	3	1	1	2	0.50	0.56	0.68 (0.55, 0.81)
27	3	2	1	3	0.19	0.56	0.25 (0.15, 0.37)
28	1	2	1	2	0.07	0.17	0.13 (0.06, 0.22)
29	1	3	2	2	0.01	0.17	0.07 (0.02, 0.16)
30	1	1	1	3	0.90	0.56	0.82 (0.71, 0.92)
31	1	1	2	3	1.00	0.95	0.86 (0.73, 0.97)
32	1	3	2	3	0.00	0.17	0.06 (0.01, 0.14)
33	1	1	1	1	0.44	0.43	0.58 (0.46, 0.7)
34	1	3	1	3	0.00	0.17	0.06 (0.01, 0.13)
35	1	2	2	3	0.75	0.56	0.47 (0.33, 0.6)
36	1	3	1	1	0.01	0.17	0.06 (0.01, 0.13)
37	2	2	1	2	0.59	0.56	0.66 (0.52, 0.78)
38	2	3	2	2	0.01	0.17	0.06 (0.01, 0.14)
39	2	1	1	3	0.59	0.95	0.92 (0.85, 0.97)
40	2	1	2	3	0.06	0.17	0.21 (0.09, 0.34)
41	2	3	2	3	0.00	0.17	0.13 (0.05, 0.24)
42	2	1	1	1	0.55	0.69	0.66 (0.54, 0.77)
43	2	3	1	3	0.60	0.56	0.48 (0.34, 0.63)
44	2	2	2	3	0.75	0.56	0.75 (0.64, 0.85)
45	2	3	1	1	0.50	0.17	0.26 (0.12, 0.42)
46	3	2	1	2	0.53	0.56	0.72 (0.61, 0.83)
47	3	3	2	2	0.98	0.95	0.89 (0.79, 0.97)
48	3	1	1	3	0.26	0.17	0.22 (0.09, 0.38)
49	3	1	2	3	0.43	0.56	0.43 (0.27, 0.59)
50	3	3	2	3	0.80	0.56	0.76 (0.65, 0.87)
51	3	1	1	1	0.50	0.56	0.82 (0.7, 0.92)
52	3	3	1	3	0.60	0.56	0.77 (0.67, 0.85)
53	3	2	2	3	0.05	0.17	0.27 (0.15, 0.41)
54	3	3	1	1	0.83	0.95	0.89 (0.78, 0.98)

Table C2

Experiment 3b. *Trial information with z-scored model predictions and participant judgments (means with 95% bootstrapped confidence intervals).* Note*: F = went fishing, T = cleared trees.*

		strength			action			full model			judgments		
index	trees	A	B	C	A	B	C	A	B	C	A	B	C
1	1	1	1	1	F	F	F	0.26	0.26	0.26	0.88 (0.69, 1.09)	0.72 (0.46, 0.92)	0.65 (0.37, 0.93)
2	1	1	1	1	T	T	F	0.41	0.41	-0.20	-0.19 (-0.51, 0.13)	-0.07 (-0.39, 0.24)	-0.96 (-1.24, -0.62)
3	1	1	1	3	F	F	T	0.46	0.46	0.57	0.79 (0.48, 1.03)	0.58 (0.21, 0.89)	0.49 (0.07, 0.91)
4	1	1	2	3	F	F	F	0.92	-0.90	-0.94	1.14 (1.00, 1.27)	-0.63 (-0.95, -0.29)	-0.70 (-1.00, -0.28)
5	1	1	2	3	T	F	T	-0.94	-0.90	0.92	-0.78 (-1.11, -0.33)	-1.10 (-1.37, -0.69)	0.87 (0.51, 1.16)
6	1	1	2	3	T	T	F	-0.94	0.88	-0.94	-1.03 (-1.23, -0.80)	0.85 (0.52, 1.16)	-1.11 (-1.35, -0.82)
7	1	1	3	3	F	F	F	0.92	-0.95	-0.95	0.83 (0.46, 1.17)	-0.21 (-0.61, 0.21)	-0.07 (-0.51, 0.38)
8	1	1	3	3	F	T	F	0.58	0.58	-0.95	0.81 (0.62, 0.98)	0.32 (0.05, 0.58)	-0.94 (-1.09, -0.78)
9	1	1	3	3	T	T	F	-0.95	0.92	-0.95	-0.90 (-1.14, -0.57)	0.95 (0.53, 1.30)	-0.81 (-1.17, -0.37)
10	1	3	1	2	F	F	F	-0.94	0.92	-0.90	-0.67 (-0.98, -0.29)	1.21 (1.05, 1.37)	-0.56 (-0.92, -0.18)
11	1	3	1	2	T	T	F	0.92	-0.94	-0.90	0.58 (0.08, 1.02)	-0.75 (-1.10, -0.34)	-0.87 (-1.04, -0.67)
12	1	3	1	3	F	F	F	-0.95	0.92	-0.95	-0.95 (-1.16, -0.64)	1.22 (1.08, 1.37)	-1.00 (-1.16, -0.83)
13	1	3	1	3	F	T	T	-0.95	-0.95	0.92	-1.17 (-1.32, -1.01)	-1.10 (-1.32, -0.83)	1.11 (0.77, 1.38)
14	1	3	3	3	F	F	F	0.05	0.05	0.05	0.53 (0.18, 0.82)	0.56 (0.25, 0.83)	0.60 (0.29, 0.85)
15	2	1	1	1	F	F	F	0.05	0.05	0.05	0.68 (0.47, 0.87)	0.69 (0.49, 0.86)	0.58 (0.35, 0.80)
16	2	1	1	1	F	T	F	0.39	-0.19	0.39	0.47 (0.19, 0.72)	-0.99 (-1.34, -0.60)	0.35 (0.06, 0.60)
17	2	1	1	1	T	T	T	0.27	0.27	0.27	-0.09 (-0.41, 0.24)	-0.02 (-0.35, 0.30)	0.01 (-0.27, 0.30)
18	2	1	1	2	F	F	F	-0.10	-0.10	0.56	-0.10 (-0.42, 0.23)	-0.12 (-0.47, 0.22)	0.65 (0.25, 0.99)
19	2	1	1	2	F	T	F	0.24	0.08	0.21	0.04 (-0.39, 0.44)	-0.55 (-0.88, -0.17)	0.52 (0.10, 0.92)
20	2	1	1	2	T	F	T	0.43	-0.10	-0.24	0.52 (0.17, 0.87)	-0.44 (-0.82, -0.01)	-0.79 (-1.08, -0.44)
21	2	1	1	3	F	F	F	0.10	0.10	-0.35	0.77 (0.60, 0.94)	0.81 (0.66, 0.96)	-0.55 (-0.88, -0.19)
22	2	1	1	3	F	F	T	-0.02	-0.02	-0.13	0.81 (0.50, 1.07)	0.84 (0.57, 1.07)	0.32 (-0.10, 0.69)
23	2	1	1	3	T	F	F	-0.47	0.44	-0.35	-1.01 (-1.24, -0.77)	0.89 (0.60, 1.15)	-0.65 (-0.97, -0.27)
24	2	1	1	3	T	F	T	-0.47	0.10	-0.02	-0.65 (-0.84, -0.45)	0.54 (0.31, 0.74)	0.47 (0.20, 0.73)
25	2	1	2	2	T	F	T	0.74	-0.02	-0.24	0.43 (0.05, 0.80)	-0.46 (-0.82, -0.07)	-0.42 (-0.80, -0.09)
26	2	1	2	3	F	F	F	-0.87	0.89	-0.89	-0.50 (-0.81, -0.17)	1.10 (0.91, 1.29)	-0.45 (-0.81, -0.06)
27	2	1	2	3	T	F	F	0.51	0.54	-0.89	-0.16 (-0.66, 0.30)	1.02 (0.86, 1.16)	-0.84 (-1.07, -0.54)
28	2	1	2	3	T	T	F	0.85	-0.91	-0.89	0.41 (-0.02, 0.80)	-0.59 (-0.95, -0.17)	-1.02 (-1.23, -0.76)
29	2	2	2	1	T	T	F	0.22	0.22	-0.77	-0.10 (-0.41, 0.18)	-0.05 (-0.36, 0.26)	-1.01 (-1.20, -0.78)
30	2	2	2	2	F	F	F	0.15	0.15	0.15	0.56 (0.36, 0.74)	0.47 (0.24, 0.68)	0.50 (0.25, 0.71)
31	2	2	2	2	T	T	F	0.51	0.51	-0.31	-0.20 (-0.49, 0.08)	0.07 (-0.26, 0.44)	-0.50 (-0.81, -0.18)
32	2	2	2	2	T	T	T	0.17	0.17	0.17	-0.12 (-0.37, 0.13)	-0.07 (-0.32, 0.17)	-0.08 (-0.35, 0.18)
33	2	2	2	3	F	F	F	0.63	0.63	-0.93	0.49 (0.17, 0.75)	0.53 (0.25, 0.75)	-0.68 (-1.03, -0.28)
34	2	3	1	2	F	F	F	-0.89	-0.87	0.89	-0.63 (-0.94, -0.27)	-0.88 (-1.18, -0.50)	1.15 (1.04, 1.26)
35	2	3	1	2	T	F	T	0.86	-0.87	-0.91	0.66 (0.17, 1.03)	-0.81 (-1.12, -0.44)	-0.71 (-1.00, -0.35)
36	3	1	1	2	F	F	F	-0.01	-0.01	0.12	0.29 (0.06, 0.50)	0.25 (-0.02, 0.50)	0.76 (0.45, 1.02)
37	3	1	1	2	F	F	T	0.33	0.33	-0.49	0.57 (0.36, 0.79)	0.56 (0.36, 0.75)	-1.07 (-1.34, -0.73)
38	3	1	1	2	F	T	F	-0.13	-0.13	0.47	-0.25 (-0.56, 0.07)	-0.92 (-1.16, -0.63)	1.04 (0.83, 1.20)
39	3	1	3	3	F	F	F	-0.84	0.45	0.45	-0.93 (-1.12, -0.68)	0.87 (0.58, 1.12)	0.88 (0.55, 1.13)
40	3	1	3	3	T	T	F	0.82	-0.24	-0.01	0.75 (0.42, 1.04)	-0.87 (-1.17, -0.50)	-0.78 (-1.04, -0.47)
41	3	2	1	2	F	F	F	0.03	0.24	0.03	0.78 (0.58, 0.95)	0.80 (0.61, 0.98)	0.66 (0.41, 0.89)
42	3	2	1	2	F	F	T	-0.09	0.58	-0.17	-0.40 (-0.63, -0.16)	0.70 (0.31, 1.04)	-1.07 (-1.22, -0.91)
43	3	2	1	2	F	T	F	0.37	-0.60	0.37	0.51 (0.21, 0.78)	-0.88 (-1.18, -0.51)	0.65 (0.37, 0.85)
44	3	2	1	2	T	F	T	-0.05	0.24	-0.05	-0.55 (-0.80, -0.28)	0.83 (0.49, 1.09)	-0.25 (-0.53, 0.06)
45	3	2	1	3	F	F	F	-0.38	-0.10	0.76	-0.62 (-0.89, -0.31)	-0.55 (-0.87, -0.16)	1.02 (0.68, 1.33)
46	3	2	1	3	F	T	F	-0.04	0.07	0.42	-0.17 (-0.61, 0.31)	-0.03 (-0.54, 0.48)	0.89 (0.50, 1.21)
47	3	2	1	3	T	F	F	0.36	0.25	0.42	-0.04 (-0.42, 0.33)	-0.31 (-0.70, 0.11)	0.67 (0.19, 1.04)
48	3	2	2	2	F	F	F	0.12	0.12	0.12	0.61 (0.45, 0.76)	0.67 (0.53, 0.80)	0.47 (0.25, 0.65)
49	3	2	2	2	F	F	T	0.46	0.46	-0.25	0.56 (0.37, 0.71)	0.44 (0.14, 0.71)	-0.92 (-1.20, -0.60)
50	3	2	2	3	F	F	F	-0.88	-0.88	0.90	-0.73 (-0.97, -0.44)	-0.79 (-1.08, -0.44)	1.18 (0.85, 1.43)
51	3	2	2	3	F	T	T	-0.88	0.86	-0.92	-0.74 (-1.11, -0.31)	0.73 (0.30, 1.11)	-0.86 (-1.20, -0.42)
52	3	2	3	3	T	T	F	0.91	-0.48	0.22	0.50 (0.10, 0.86)	-0.61 (-1.00, -0.19)	-0.41 (-0.84, -0.02)
53	3	3	1	2	T	F	T	-0.44	-0.10	0.70	-0.92 (-1.07, -0.76)	-0.82 (-0.99, -0.64)	0.69 (0.50, 0.88)
54	3	3	1	2	T	T	F	-0.44	0.42	-0.38	-0.88 (-1.24, -0.37)	0.64 (0.28, 0.94)	-1.10 (-1.29, -0.89)
55	3	3	2	2	F	F	F	0.90	-0.88	-0.88	1.08 (0.75, 1.30)	-0.68 (-0.95, -0.39)	-0.66 (-0.94, -0.34)
56	3	3	2	2	F	T	T	0.44	0.40	0.40	0.97 (0.62, 1.21)	0.28 (-0.11, 0.66)	0.30 (-0.07, 0.69)
57	3	3	2	3	F	F	F	0.68	-0.93	0.68	0.39 (0.00, 0.71)	-0.82 (-1.08, -0.50)	0.55 (0.28, 0.81)
58	3	3	3	1	T	T	F	0.22	0.22	-0.84	-0.18 (-0.55, 0.20)	0.12 (-0.24, 0.47)	-0.93 (-1.21, -0.60)
59	3	3	3	3	T	T	F	0.62	0.62	-0.41	-0.14 (-0.45, 0.10)	0.13 (-0.08, 0.35)	-0.90 (-1.20, -0.57)
60	3	3	3	3	T	T	T	0.27	0.27	0.27	0.36 (0.04, 0.66)	0.35 (-0.05, 0.66)	0.40 (0.10, 0.67)

Table C3

Experiment 3c*. Trial information. The missing values (indicated by '–') had to be inferred by participants. For example, in trial 1, participants had to infer the number of trees. In trial 32, they had to infer what actions fisherman B and C took.* Note*: F = went fishing, T = cleared trees.*

		strength			action			praise		
trial	trees	A	B	C	A	B	C	A	B	C
1	–	1	1	3	F	F	F	high	high	low
2	–	3	1	1	F	F	F	low	medium	medium
3	–	1	3	1	F	F	F	low	high	low
4	–	1	1	3	T	T	T	medium	medium	high
5	–	3	1	1	T	T	T	high	low	low
6	–	1	3	1	T	T	T	high	low	high
7	2	–	2	3	T	T	F	high	low	low
8	2	3	–	2	F	T	T	low	medium	medium
9	2	3	3	–	T	F	T	medium	medium	high
10	2	3	–	3	T	T	F	high	low	low
11	2	3	3	–	F	T	T	low	high	high
12	1	1	3	2	–	T	F	low	high	low
13	1	1	2	3	–	F	T	high	low	high
14	3	3	2	2	F	–	F	high	medium	low
15	3	2	3	2	F	F	–	low	high	low
16	3	3	1	2	F	–	T	medium	medium	medium
17	3	2	3	1	T	F	–	medium	medium	high
18	–	1	2	1	–	F	–	high	low	high
19	3	–	–	3	T	T	T	medium	medium	low
20	3	1	–	1	F	–	F	low	high	low
21	3	2	1	–	F	F	F	low	low	high
22	3	1	1	2	–	–	F	low	low	high
23	–	1	2	2	T	T	T	low	medium	medium
24	–	2	1	2	T	T	T	medium	medium	medium
25	–	2	2	1	T	T	T	high	high	low
26	1	2	3	1	T	–	T	high	high	low
27	1	2	1	1	–	F	F	high	medium	medium
28	1	1	2	1	F	–	F	medium	low	medium
29	3	1	2	3	F	–	–	high	medium	medium
30	3	3	1	2	–	F	–	high	medium	medium
31	2	1	1	2	–	–	F	high	medium	medium
32	2	2	1	1	F	–	–	high	medium	medium
33	3	1	3	–	T	F	F	medium	medium	low
34	3	1	–	3	T	F	F	medium	high	medium
35	3	3	1	–	F	T	F	medium	medium	medium
36	3	3	3	1	–	T	F	high	high	low