A computational framework for understanding the roles of simplicity and rational support in people's behavior explanations

Alan Jern
Department of Humanities, Social Sciences, and the Arts
Rose-Hulman Institute of Technology

Austin Derrow-Pinion
Department of Computer Science and Software Engineering
Rose-Hulman Institute of Technology

AJ Piergiovanni
School of Informatics and Computing
Indiana University

Abstract

When explaining other people's behavior, people generally find some explanations more satisfying than others. We propose that people judge behavior explanations based on two computational principles: simplicity and rational support—the extent to which an explanation makes the behavior "make sense" under the assumption that the person is a rational agent. Furthermore, we present a computational framework based on decision networks that can formalize both of these principles. We tested this account in a series of experiments in which subjects rated or generated explanations for other people's behavior. In Experiments 1 and 2, the explanations varied in what the other person liked and disliked. In Experiment 3, the explanations varied in what the other person knew or believed. Results from Experiments 1 and 2 supported the idea that people rely on both simplicity and rational support. However, Experiment 3 suggested that subjects rely only on rational support when judging explanations of people's behavior that vary in what someone knew.

Keywords: behavior explanation, mental state inference, theory of mind, simplicity, probabilistic model, decision networks

---

Corresponding author: Alan Jern, Department of Humanities, Social Sciences, and the Arts, Rose-Hulman Institute of Technology, 5500 Wabash Ave, Terre Haute, IN 47803. E-mail: jern@rose-hulman.edu.
Materials, code, and data are available at `https://github.com/alanjern/behavior-explanation`.

Introduction

People are constantly trying to explain other people's behavior. Whenever you wonder about the motives, beliefs, desires, or influences behind someone else's behavior—that is, whenever you ask yourself why someone did something—you are attempting to explain that person's behavior.

As an example, suppose you saw a reckless driver careening down the road, far past the speed limit. You might wonder why she was doing that; that is, you might want to explain her behavior. Consider three possible explanations: (1) she was late to a meeting and was in a hurry; (2) she did not know the speed limit; and (3) her speedometer was broken and she did not know how fast she was going. Different people might disagree about how satisfying these explanations are, but people will likely agree that all of these explanations are more satisfying than the following explanation: she was speeding because she was late to a meeting, *and* she did not know the speed limit, *and* her speedometer was broken. Additionally, there are some explanations that most people will not find satisfying at all. For example, consider this explanation: she was speeding because she knew that George Washington was the first president of the United States.

How do people judge behavior explanations, and what makes some behavior explanations more satisfying than others? These questions are related to the broad study of social cognition and how people make judgments about other people's behavior (Kelley & Michela, 1980; Gilbert, 1995). Specifically, behavior explanation is closely related to interpersonal attribution: how people decide whether to attribute other people's behavior to either their individual characteristics or to situational factors. Some early research on interpersonal attribution focused on normative inference principles for making judgments about other people's behavior (Jones & Davis, 1965; Kelley, 1973). But most work on interpersonal attribution, particularly of the past few decades, has focused on the cognitive processes that drive these judgments, rather than the computational principles that underlie them (Anderson, Krull, & Weiner, 1996; Gilbert, 1998). The lack of focus on computational principles has meant that the majority of theories that have emerged from this research have been qualitative and are not capable of making quantitative predictions (Korman, Voiklis, & Malle, 2015). Consequently, our understanding of how people explain behavior is broad but shallow.

In this paper, we propose two computational principles that underlie people's judgments about behavior explanations. We call the two principles *simplicity* and *rational support*. We show that both principles can be formally instantiated using the computational framework of decision networks. This paper therefore represents a first step toward formalizing qualitative theories of behavior explanation (Korman et al., 2015).

Later, we provide a formal definition of the two principles. For now, we will provide a brief informal introduction to each principle using the reckless driver example from earlier. The simplicity principle states that simpler behavior explanations are more satisfying. For example, consider the explanation for the driver's behavior that gives three separate reasons for speeding (she was late to a meeting, she did not know the speed limit, and her speedometer was broken). This explanation does explain the driver's behavior, but it seems to *over*-explain or overdetermine her behavior. That is, it gives three independently sufficient reasons when each one would suffice. By comparison, each one of the three component

explanations is simpler and therefore seems more satisfying than the full explanation.

The rational support principle states that a behavior explanation will be more satisfying to the extent that it makes the behavior "make sense", under the assumption that the behavior was carried out by a rational agent. For example, consider the explanation for the driver's behavior that she was speeding because she knew that George Washington was the first president of the United States. This knowledge provides no rational support for speeding—it does not motivate someone to be in a hurry or to drive recklessly. As a result, this "explanation" does not explain anything at all. By contrast, someone who is late to a meeting would likely want to hurry to get to the meeting as quickly as possible. Therefore, explaining the driver's speeding by referring to the fact that she was late to a meeting does provide rational support for her behavior, making the explanation that mentions her being late more satisfying.

Both of these principles have existing empirical support. Evidence for the simplicity principle comes from work on causal explanation by psychologists (Keil & Wilson, 2000; Lombrozo, 2006, 2012; Legare & Clegg, 2015; Pacer, Williams, Chen, Lombrozo, & Griffiths, 2013; Pacer & Lombrozo, 2017) and machine learning researchers (Pacer et al., 2013; Flores, Gámez, & Moral, 2005; Yuan & Lu, 2007; Nielsen, Pellet, & Elisseeff, 2008; DeJong, 2006). For example, in one study (Lombrozo, 2007), subjects learned about several diseases that caused overlapping symptoms. Then, after reading about a patient's symptoms, subjects rated explanations for the patient's symptoms such as "the patient has diseases X and Y". In general, subjects rated simpler explanations (explanations that invoked fewer causes) to be "better and more likely to be true" (Lombrozo, 2007). While psychological research on how people generate and judge causal explanations has been extensive, research on behavior explanation has been more limited. One relevant line of research has been motivated by Kelley's (1987) discounting principle, which is similar to our simplicity principle. The discounting principle states that people will discount the strength of a cause when alternative causes are present. The discounting principle has been taken by interpersonal attribution researchers to suggest that people will prefer behavior explanations with fewer causes. Specifically, studies showing that people sometimes rate explanations of other people's behavior that invoke more reasons as better than explanations that invoke fewer reasons (e.g., Leddo, Abelson, & Gross, 1984) have been presented as evidence against the discounting principle (McClure, 1998; Morris, Smith, & Turner, 1998).

However, behavior explanations likely rely on more than simplicity or discounting because behavior explanations are different than causal explanations in several respects. First, behavior explanations are more likely to refer to mental states like beliefs and desires that gave rise to an intention to act (Malle, 1999). Second, when reasoning about other people's mental states, people expect others to behave in a goal-directed way (Baker, Saxe, & Tenenbaum, 2009; Ullman et al., 2010; Baker & Tenenbaum, 2014; Jern & Kemp, 2015; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017), an expectation that can even influence their causal judgments (Lagnado & Channon, 2008; Kirfel & Lagnado, 2019). This observation leads to the rational support principle.

Some evidence for the rational support principle in behavior explanation comes from work by Malle (1999, 2004), who has proposed the most complete psychological theory to date of behavior explanation. One claim of this theory is that explanations of intentional

behavior are based on *reasons* rather than *causes*. Malle defines reasons as "agents' mental states whose content they considered and in light of which they formed an intention to act" (Malle, 1999). By contrast, explanations of unintentional behaviors like blushing or sweating are more likely to be based on simple causes that do not qualify as reasons. For a reason-based explanation to be considered valid, it must provide rational support for the behavior, meaning that the beliefs and desires identified in the explanation must make the behavior rational. Malle's theory has considerable empirical support (Malle, 1999; Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000; O'Laughlin & Malle, 2002; Malle, Knobe, & Nelson, 2007), but the theory is limited by its qualitative nature. For example, the theory can predict whether someone will generate a reason explanation or a cause explanation for a given behavior, but cannot predict how probable it is someone will generate a particular reason explanation from a set of possibilities. Malle and his colleagues themselves have stated that more computational work is needed (Malle, 2004; Korman et al., 2015).

Computational framework

In this section, we describe the computational framework of decision nets (short for decision networks[1]), and explain how decision nets can be used to formally instantiate the simplicity and rational support principles. Previous research has suggested that people represent other people's choices and behavior using decision nets (Jern & Kemp, 2015; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015) or similar representations, like Markov decision processes (Baker et al., 2009; Baker & Tenenbaum, 2014; Baker et al., 2017; Tauber & Steyvers, 2011). We propose that people use representations like decision nets to judge explanations for others' behavior. Specifically, we propose that each explanation is represented by a differently structured decision net and that people judge explanations by determining which decision net structure best captures an observed behavior.

It is important to keep in mind that our computational framework assumes that decision nets are similar to the representations of a situation *from an observer's perspective*. As a result, decision net explanations may not accurately represent the objectively correct explanations for someone's behavior if an observer has false or incomplete information about a situation.

Decision nets are related to causal Bayes nets (Gopnik et al., 2004; Sloman, 2005). Both decision nets and causal Bayes nets can be depicted as graphs in which the edges or connections indicate causal relationships between variables (see Figure 1b for an example). The main difference between decision nets and Bayes nets is that decision nets treat intentional choices differently than other variables. Specifically, decision nets assume that choices are made to increase the decision maker's expected utility, whereas Bayes nets have no built-in notion of expected utility. For this reason, decision nets are well suited for modeling how people reason about other people's behavior and have been used previously as psychological models of social cognition. Specifically, they have been used to model mental state inference (Jern & Kemp, 2015) and moral decision-making (Kleiman-Weiner et al., 2015).

We will introduce the decision net framework with an example based on the story that we used in Experiment 1. Suppose Lori is attending a meeting. There are some people

[1]Decision networks are also known as influence diagrams (Howard & Matheson, 2005).

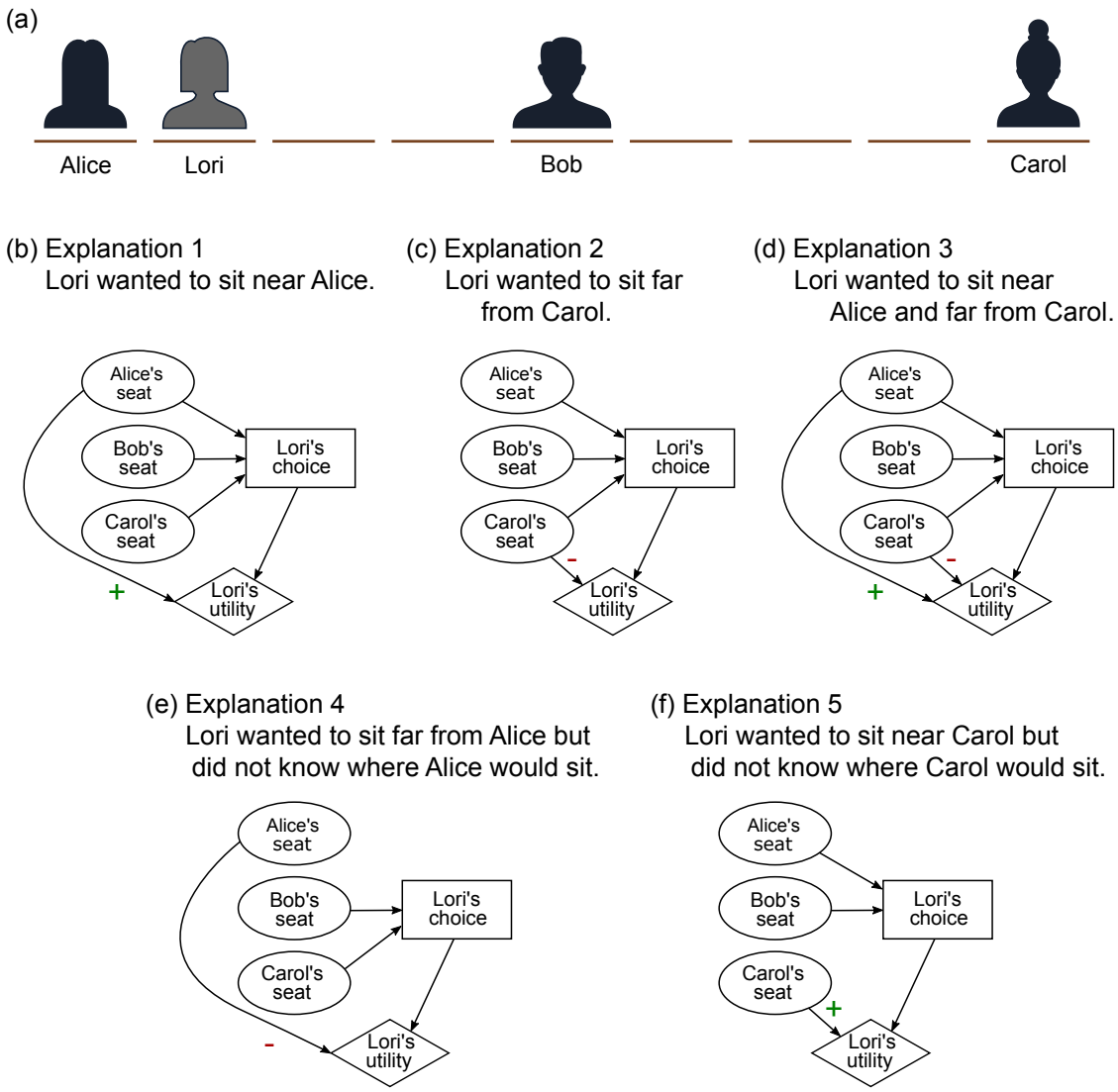

Figure 1. An example behavior and several decision net explanations for the behavior. (a) Why did Lori choose to sit where she did? (b)–(f) Different possible decision net representations of explanations for Lori's behavior. Explanations 1–3 vary only in who Lori likes and dislikes. Explanations 4 and 5 also vary in what Lori knew when she arrived at the meeting. The plus and minus signs indicate whether Lori likes (+) or dislikes (−) that person.

at the meeting that Lori likes, some people that she dislikes, and some people that she is indifferent toward. The seats for the meeting are arranged in a single row. Before Lori arrives, three people—Alice, Bob, and Carol—are already seated. Lori must choose a seat from the remaining open seats. This situation is depicted in Figure 1a. The figure shows where the four people ended up sitting.

This situation can be represented by the decision net in Figure 1b. There are four main variables in the graph: Alice's, Bob's, and Carol's seat locations, and Lori's choice about where to sit. From Lori's standpoint, the other people's seat choices are existing states of the world; decision nets represent state-of-the-world variables using oval-shaped nodes. As we previously noted, decision nets treat intentional choices, like Lori's choice about where to sit, differently than other variables. Thus, Lori's choice is represented using a rectangular node. Finally, decision nets include a utility variable, represented by a diamond-shaped node, for each decision maker represented in the graph.

The edges in the decision net represent relationships between variables. In the decision net in Figure 1b, the edges leading from the oval nodes to Lori's choice node indicate that Lori knows where Alice, Bob, and Carol are seated before she makes her choice. If Lori had arrived before Alice, the edge leading from Alice's location to Lori's choice would be absent. Edges leading to the utility node indicate what variables contribute to the decision maker's utility. In Figure 1b, there are edges leading from the node for Alice's seat and for Lori's choice. This indicates that Lori's utility is a function of Alice's location and Lori's location but not of Bob's and Carol's locations. For a more detailed introduction to using decision nets as psychological models, see Jern and Kemp (2015).

Now suppose you see Lori sit in the seat closest to Alice, as shown in Figure 1a. Why did Lori choose to sit there? Just like with the reckless driver, there are multiple possible explanations. Each explanation can be represented by a different decision net. We will consider five explanations, represented by the decision nets in Figures 1b–f. The first three explanations in the figure assume that Lori knew where all three of her coworkers were sitting before she chose where to sit:

- **Explanation 1** (Figure 1b): Lori sat there because she wanted to sit near Alice.
- **Explanation 2** (Figure 1c): Lori sat there because she wanted to sit far from Carol.
- **Explanation 3** (Figure 1d): Lori sat there because she wanted to sit near Alice and far from Carol.

Because Lori ended up sitting as close as possible to Alice and as far as possible from Carol, these three explanations have high rational support. Explanation 3 is just the conjunction of Explanations 1 and 2. Therefore, Explanations 1 and 2 are simpler than Explanation 3. In the next section, we will elaborate on how decision nets formalize these statements.

Explanations 1–3 all assume that Lori knew where her coworkers were sitting. Therefore, the decision nets for these explanations all have edges leading from the nodes representing Lori's coworker's locations to the node representing her choice. The explanations differ in which people Lori cared about. Therefore, the decision nets differ in which edges lead to the utility node. Additionally, each of the edges leading to the utility node on each decision net have a plus or minus sign indicating the effect that the person's location has on

Lori's utility. A plus sign (+) indicates that Lori is happier the closer she is to that person; a minus sign (−) indicates that Lori is less happy the closer she is to them.

The next two explanations do not assume that Lori knew where all three coworkers were sitting before she chose where to sit:

- **Explanation 4** (Figure 1e): Lori sat there because she wanted to sit far from Alice but did not know where Alice would sit.
- **Explanation 5** (Figure 1f): Lori sat there because she wanted to sit near Carol but did not know where Carol would sit.

Explanation 4 also provides rational support because if Lori wanted to sit far from Alice and *did* know where Alice was going to sit, Lori would have chosen a different seat. In other words, Lori's behavior is only rational if she did not know where Alice was going to sit. Similarly, under Explanation 5, if Lori had known where Carol was going to sit, she would have chosen a seat closer to Carol. Therefore, Lori's lack of knowledge about where Carol would sit helps to provide a satisfying explanation for her behavior.

Suppose you were asked to judge which explanations for Lori's behavior from this set of five are most satisfying. This is what we asked subjects to do in our experiments. Recall that we hypothesized that people would base their judgments on how simple an explanation was and how much rational support it provided. We now show how simplicity and rational support can be formally instantiated using the decision nets in Figure 1.

Simplicity

Because decision nets are networks, we may quantify how simple a decision net explanation is using standard methods of measuring network complexity. Intuitively, simpler networks have fewer nodes, edges, and possible node values. This intuition can be captured using a definition of simplicity based on minimum description length (Rissanen, 1978). Minimum description length has been used before to account for aspects of reasoning (Fass & Feldman, 2002). For example, people find it easier to learn concepts that can be described by shorter codes (Feldman, 2000).

Let $S(N)$ be the simplicity of decision net N. We define $S(N)$ as the inverse of a standard minimum description length-based definition of network complexity (de Campos, 2006):

$$S(N) = \frac{1}{\sum_i (x_i \cdot q_i)}, \tag{1}$$

where x_i is the number of values that node i can take on, and q_i is number of values that the parents of i can take on. According to this definition, simplicity increases as the number of nodes, edges, and possible values of each node decreases.

There are other ways to quantify simplicity. For example, Pacer and Lombrozo (2017) draw a distinction between count simplicity and root simplicity for causal explanations. Conceptually, count simplicity measures the total number of causes invoked in an explanation and root simplicity measures just the number of *root* causes—rather than proximate causes[2]—to explain something. In the examples used so far, these two types of simplic-

[2]Suppose someone coughs. The proximate cause of the cough was mucus buildup in the lungs but the root cause was a cold.

ity are equivalent because all of the factors identified in Explanations 1–5 are root causes. That is, they do not have their own explanations. To illustrate the difference between count and root simplicity, suppose that the people at the meeting in the example above are all members of competing teams of two and that the teams dislike each other. Now you might explain Lori's choice of seat by saying that she and Alice are team members. This fact would provide a single root cause for Lori liking Alice and Lori disliking Carol. As a result, this explanation would have higher root simplicity than Explanation 5, which says that Lori separately likes Alice and dislikes Carol, thus providing two root causes. Because all of the explanations in our experiments involved only root causes and therefore have equal count and root simplicity, we will ignore the distinction between count and root simplicity for most of this paper, but we revisit it in the General Discussion.

Rational support

Decision nets assume that each choice c is taken to increase the decision maker's utility. Formally, let $EU(c_i)$ be a decision maker's expected utility from choosing option c_i. A decision function specifies how decision makers choose among their options as a function of expected utilities. A natural choice for a decision function is a utility-maximizing function, which specifies that a decision maker will always choose the option c_i that has the highest expected utility:

$$P(c = c_i) = \begin{cases} 1, & \text{if } c_i = \arg\max_c EU(c) \\ 0, & \text{otherwise} \end{cases}.$$

However, people do not always maximize utility (Vulkan, 2002). Additionally, people's choices and behavior are usually partly influenced by factors that are unknown to observers. Consistent with these observations, studies have found that people's inferences about other people's choices are better explained by a probabilistic decision function than a maximizing decision function (Jern & Kemp, 2015; Jern, Lucas, & Kemp, 2017; Lucas et al., 2014). For example, a decision function might specify that a decision maker chooses options probabilistically in proportion to their expected utilities (i.e., a softmax function):

$$P(c = c_i) = \frac{EU(c_i)}{\sum_j EU(c_j)}. \tag{2}$$

The decision function of a decision net, regardless of its specific functional form, provides a way to formalize the principle of rational support. We propose that explanations (i.e., decision nets) that result in the observed choice having higher probability according to the decision net's decision function have higher rational support. To illustrate, suppose that Lori knew where Alice, Bob, and Carol were sitting before Lori chose where to sit. She chose the seat closest to Alice and furthest from Carol, as shown in Figure 1a. As explained above, Explanations 1–3 all provide rational support because this is the seat that would provide Lori with the most utility if she likes Alice, dislikes Carol, or both. By contrast, consider the following explanation:

- **Explanation 6**: Lori sat there because she dislikes Alice.

This explanation provides little rational support because if Lori dislikes Alice, her choice of seat would provide her little utility (perhaps negative utility). Thus, it is a very poor explanation.

Comparing explanations

Sometimes the simplicity and rational support principles will be in conflict. For example, one explanation might be simple but provide weak rational support, and one explanation might be complex but provide strong rational support. We propose that when people are comparing explanations for a behavior, they reconcile the two principles through probabilistic model selection in which people have a bias in favor of simpler explanations.

Formally, we compute the probability that each decision net explanation N is the explanation for choice c as follows:

$$P(N|c) \propto P(c|N) \cdot P(N). \tag{3}$$

The prior probability, $P(N)$, is proportional to the simplicity of the decision net: $P(N) \propto S(N)$. This captures the idea that simpler explanations will receive *a priori* greater probability. The likelihood, $P(c|N)$, is the decision function of the decision net. This captures the rational support principle. Explanations that provide stronger rational support will result in the probability $P(c|N)$ being higher and therefore receiving higher posterior probability. This formalization can handle cases in which simplicity and rational support are in conflict. A complex explanation N will have a low prior probability $P(N)$. But if the explanation provides strong rational support (i.e., $P(c|N) \approx 1$), the posterior probability $P(N|c)$ might still be higher than alternative simpler explanations with weak rational support.

Overview of experiments

We predicted that people would rate explanations with higher posterior probability according to Equation 3 as more satisfying. We tested this prediction in a series of experiments in which people read descriptions of people's behavior and then rated explanations for why the people did what they did. Experiments 1 and 2 examined explanations that differed in what a person liked or disliked, as in Explanations 1–3 for Lori's behavior. Experiment 3 examined explanations that differed in what a person knew, as in Explanations 4 and 5 for Lori's behavior.

To preview our findings, the results of Experiment 1 indicated that subjects did incorporate both simplicity and rational support into their judgments but not in the way predicted by our decision net model in Equation 3. This led us to consider an alternative model that combined both principles in a non-probabilistic way and that predicted subjects' judgments much better. Experiment 3 indicated that subjects incorporated simplicity but not rational support into their judgments, a result we investigated further and confirmed in a follow-up experiment (Experiment 4 in the Supplementary Materials). We discuss possible explanations for the different results of Experiments 1 and 3 in the Discussion for Experiment 3 and in the General Discussion.

Experiment 1A: Judging behavior explanations that differ in what someone likes and dislikes

We used the same story in Experiments 1A and 1B as in our running example involving Lori's choice of seat at a meeting. Subjects read about someone who arrived last at a meeting in which there was a single row of chairs and chose a seat. Subjects then rated explanations for why the person chose that seat.

Near A	Near A and B	Near A and B, but far from C
Near B	Near A, and far from B	Near A, but far from B and C
Far from A	Near A, and far from C	
Far from B	Near B, and far from C	
Far from C	Far from A and C	
	Far from B and C	

Table 1
The set of explanations used in Experiments 1A and 1B. The letters A, B, and C were replaced with names. For clarity and consistency, A, B, and C here correspond to Alice, Bob, and Carol in Figure 3.

Method

Participants. 105 subjects completed the experiment on Amazon Mechanical Turk. Data from 20 additional subjects were omitted because the subjects failed an attention check described below. All subjects were paid.

Design and Procedure. Subjects were randomly assigned to one of three conditions. The conditions are depicted in Figure 3. In each condition, subjects saw a diagram like the ones in the figure and were told: "Alice, Bob, and Carol are attending a meeting and sat in the seats marked below. Lori showed up to the meeting next. In this meeting, there are some people Lori likes, some she dislikes and some she is indifferent toward. [Diagram appeared here.] Lori sat in the seat marked above." The names of the four people varied randomly between trials.

Subjects were then instructed to "rate how satisfying the following explanations are for why Lori chose to sit there." They saw 13 explanations, all of which were expressed as combinations of preferences to sit near to, or far from, certain people. Table 1 shows the complete set of explanations. The explanations in the table are expressed in shorthand. For example, the second explanation in the second column of Table 1, "Near A, and far from B", was presented to subjects as "Lori wanted to sit near Alice and far from Bob." The set of explanations includes any explanation that would provide complete rational support for the person's seat choice in any of the three conditions, as well as several explanations, such as "Far from A", that do not provide strong rational support in any condition. Subjects rated each explanation on a scale from 1 ("Very bad explanation") to 7 ("Very good explanation").

Subjects saw all 13 explanations at once on a single screen with the order randomized for each subject. We did this to make the rating task as similar as possible to the method by which our model compares explanations. Specifically, because the probabilistic predictions of the model are made relative to one another, it was important that subjects knew the full set of possible explanations when making their judgments.

Half of subjects in each condition saw a mirror image of the diagrams in Figure 3. For example, in Condition 1, the person identified by X would be shown seated closest to Person C instead of Person A. We adjusted the explanations for these subjects to reflect the different seating locations. For example, the phrase "Near A" in the explanations was replaced with "Near C".

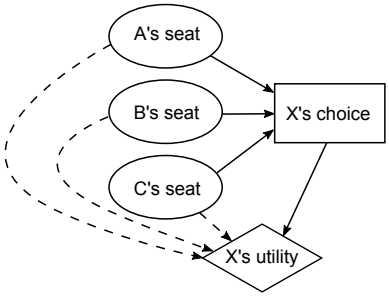

Figure 2. The decision net model used to represent the explanations in Experiments 1A and 1B. The explanations in Table 1 differed in the presence or absence of the dashed edges.

Attention check. After subjects submitted their ratings, we included an attention check. The attention check consisted of a screen that looked identical to the previous one (a diagram like the ones in Figure 3 with a list of explanations) but with different instructions: "This case is only meant to check if you are reading the instructions carefully. Please ignore the information and leave all the questions blank." If a subject failed to follow these instructions, all of their data were excluded from analysis.

Model

Each of the explanations in Table 1 can be represented by its own decision net. Figure 2 shows a template decision net upon which all the explanations in Table 1 are based. Because the person whose behavior is being explained (e.g., Lori) arrived last, she knew where everyone else was seated before she made her choice. As a result, every decision net explanation includes edges leading from the seat locations for A, B, and C, to Person X. The explanations do differ in which people Person X cares about, captured by the presence or absence of edges leading from the seat location nodes to Person X's utility node. The dashed edges in Figure 2 indicate that these edges were present in only some of the explanations. For example, in the decision net corresponding to the "Near A, and far from B" explanation, only the "A's seat" and "B's seat" nodes would have edges leading to the utility node.

Additionally, decision net explanations with identical network structures, such as the "Near A" and "Far from A" explanations, are captured by differences in their utility functions (previously represented using plus and minus signs in Figure 1). We assumed that the total utility $U(s)$ that Person X assigned to each seat s depended on the seat's proximity to the people X wanted to sit near to or far from. Specifically, let $u_i(s)$ be the utility X derives from seat s's proximity to Person i. We defined $u_i(s)$ as follows:

$$u_i(s) = \begin{cases} e^{-kd} & \text{if X wants to sit near } i \\ 1 - e^{-kd} & \text{if X wants to sit far from } i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

In this equation, d is the distance in number of seats from Person i's seat, and k is a free parameter that affects how quickly X's utility changes as a function of distance from Person

i. Equation 4 has the property that the change in utility Person X gets as they move farther away from Person i is much greater when they are close to i. The intuitive rationale behind this choice of utility function can be illustrated with the following example. If you are friends with someone, you get the most benefit from sitting immediately next to them. You would get considerably less benefit if you were two seats away from them because it would be harder to talk. After a certain distance, it would hardly matter how far away you were because being one more seat away would not make it any harder to talk. This reasoning applies in reverse for someone you dislike and want to sit far from.[3]

Because some of the explanations include references to multiple people, we made a standard assumption that utilities are independent and additive. That is, we assumed that Person X's total utility $U(s) = \sum_i u_i(s)$.

Alternative models. In order to test our prediction that people rely on both simplicity and rational support, we also considered two versions of the model that each omitted one of these principles.

Simplicity-only model. The simplicity-only model does not take rational support into account and only assumes that explanations are better to the extent that they are simpler. Formally, the model compares explanations as follows:

$$P(N|c) \propto P(N). \tag{5}$$

The simplicity-only model is equivalent to setting $P(c|N) \propto 1$ in the full model, meaning that the model assumes that all choices are equally probable regardless of the explanation.

Rational support-only model. The rational support-only model does not take simplicity into account and only assumes that explanations are better to the extent that they provide more rational support for the choice. Formally, the model compares explanations as follows:

$$P(N|c) \propto P(c|N). \tag{6}$$

The rational support-only model is equivalent to setting $P(N) \propto 1$ in the full model, meaning that it places a uniform prior probability on all explanations.

Results

Model predictions. Model predictions were generated according to Equations 3, 5, and 6 for all values of k from 0.05 to 3 in increments of 0.05. We then normalized each set of predictions so that the probabilities for each condition summed to 1. We determined the best-fitting value of k for each model using the following procedure: For each setting of k, we used a 10-fold cross-validation training scheme, using each training set to fit a linear regression model predicting subjects' mean ratings from the model's predictions, and computing the average root mean squared error (RMSE) of the validation sets. We then chose the value of k that minimized RMSE. We used this same procedure to fit model parameters throughout the paper.

[3] One consequence of the utility function in Equation 4 is that the utility that Person X derives from someone they want to sit far from is always positive, meaning that X would receive more positive utility from people they want to sit far from, no matter how close those people are, than people X is indifferent to. This might seem unusual, and one alternative would be to set X's utility $u_i(s) = -e^{-kd}$ for people that X wants to sit far from. However, as we show in Appendix A, the choice between these two options for the utility function has a minimal effect on overall model performance.

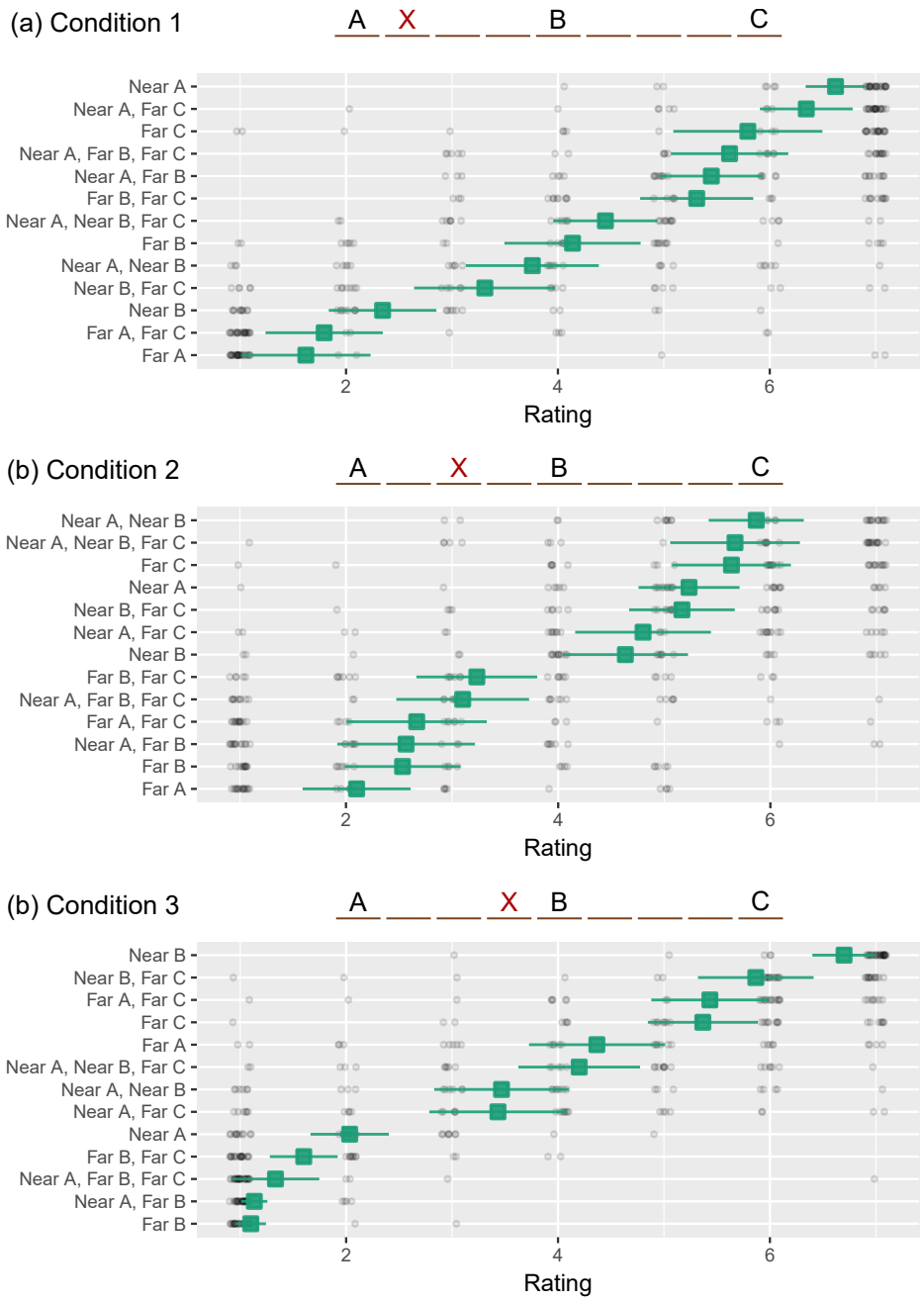

Figure 3. Subjects' mean ratings for how satisfying each explanation was for each condition in Experiment 1A. In each condition, the explanations are ordered by mean rating. The error bars are 95% confidence intervals. The black points indicate individual subjects' responses. Some horizontal jitter has been added to these points to make them more easily visible.

Subject ratings. Figure 3 shows subjects' mean ratings for all explanations in all conditions. In the figure, the explanations in each condition are ordered by mean subject rating. Subjects' judgments were reasonable overall. For example, in Condition 1 (Figure 3a), the three highest-rated explanations were "X wanted to sit near A", "X wanted to sit near A, and far from C", and "X wanted to sit far from C". By contrast, none of these explanations were in the top three highest-rated explanations in Condition 3.

We analyzed the data by fitting a Bayesian ordinal regression model using the R `brms` package (Bürkner, 2017). To test our general prediction that subjects would take both rational support and simplicity into account when making their judgments, we fit a model using subject rating as the dependent variable with the best-fitting rational support-only model predictions and simplicity-only model predictions (and their interactions) as population-level predictors, and fit separate intercepts for each subject[4]. We placed a positive but diffuse prior probability distribution of $\mathcal{N}(5, 10)$ on population-level coefficients to reflect the prior research indicating that both factors would have some effect on subjects' ratings.

The estimated coefficient for the rational support effect was 18.6, 95% CI $[17.0, 20.2]$; the estimated coefficient for the simplicity effect was 26.0, 95% CI $[8.7, 11.6]$. The credible intervals for both coefficients did not include 0, providing evidence that the simplicity and rational support values were both predictive of subjects' ratings, therefore suggesting that subjects took both factors into account.

Comparison between model predictions and subject ratings. Figure 4 compares subjects' mean ratings with the best-fitting model predictions for the full decision net model and the two alternative models. Although the models were fit using RMSE, we report correlation coefficients here for ease of comparison. The simplicity-only model performed very poorly ($r = 0.004$). This is perhaps not surprising considering that this model essentially only takes into account the number of factors referenced in each explanation. For example, in Condition 1, in which Person X sat closest to Person A, the simplicity-only model assigns equal probabilities to the explanation "X wanted to sit near A" and "X wanted to sit far from A". By contrast, as shown in Figure 3a, these two explanations were, respectively, the highest-rated and lowest-rated overall by subjects. The poor performance of the simplicity-only model strongly suggests that, consistent with the regression analysis, subjects took more than simplicity into account.

But contrary to predictions, the full decision net model did not predict subjects' ratings very well ($r = 0.48$, best-fitting $k = 0.3$). And, in fact, the rational support-only model predicted subjects' ratings better overall ($r = 0.73$, best-fitting $k = 0.25$) than the full decision net model. These results not only provide evidence against the decision net model but appear to suggest that people are not placing much emphasis on simplicity in their judgments, despite the fact that simplicity had a predictive role in the regression analysis.

The role of simplicity in subjects' ratings

Examining subjects' ratings that deviated the most from the decision net model predictions in Figure 4 supports the hypothesis that subjects placed less emphasis on simplicity

[4]In R model syntax, we fit the model: `rating ~ rational_support * simplicity + (1|subject)`.

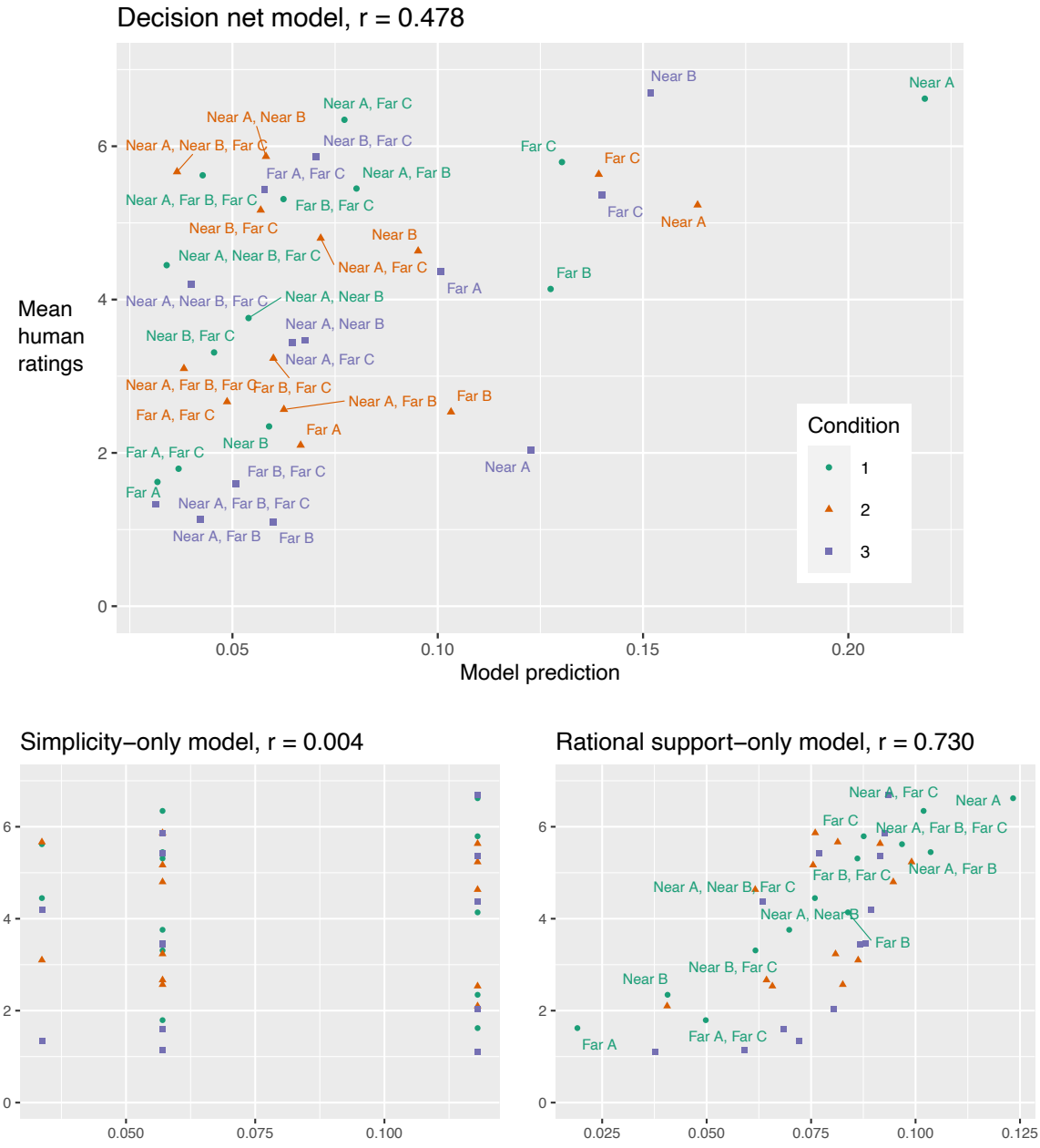

Figure 4. Comparison between mean subject ratings and model predictions for all three models in Experiment 1A. Each point corresponds to one explanation from one condition. The x-axes indicate each model's predicted probabilities. For the rational-support model, only the explanation labels for Condition 1 are included for readability.

than the decision net model. For example, the "Near A, Far B, Far C" explanation was rated higher by subjects than the decision net model in Conditions 1 and 2. Similarly, the "Near A" rating was rated much lower by subjects than the decision net model in Condition 3. All three outcomes are consistent with subjects placing less emphasis on simplicity than the model.

But subjects did not ignore simplicity. Consider the mean ratings for Condition 1 in Figure 3a. While the explanations "X wanted to sit near A" and "X wanted to sit near A, but far from B and C" are both consistent with X's choice (i.e., X's choice is the optimal choice under both explanations), subjects gave lower mean ratings to the second, less simple explanation. This difference in ratings was statistically significant, according to a one-tailed paired-sample t-test, $t(28) = 3.68, p < 0.001$ (difference in means of 1.00, Cohen's $d = 0.79$). Similarly, in Condition 3 in Figure 3b, the explanations "X wanted to sit near B" and "X wanted to sit far from A and C" are both consistent with X's choice, but subjects gave lower mean ratings to the second, less simple explanation. This difference in ratings was statistically significant, according to a one-tailed paired-sample t-test, $t(29) = 4.68, p < 0.001$ (difference in means of 1.27, Cohen's $d = 0.98$).

However, as Liddell and Kruschke (2018) note, analyzing ordinal data, like subjects' ratings, with metric models like t-tests may lead to statistical errors. Therefore we fit a second Bayesian ordinal regression model using subject rating as the dependent variable with condition (1–3) and explanation (from Table 1) (and their interactions) as population-level predictors, and fit separate intercepts for each subject[5]. After fitting the model, we performed hypothesis tests comparing the estimated coefficients for the interaction terms assigned to two pairs of explanations used in the t-tests.

Contrary to the results of the t-tests, our data were only slightly more likely to be observed ($BF_{10} = 0.39$) if the estimated interaction coefficient for the Condition 1 "Near A" explanation (4.31, 95% CI $[2.75, 5.94]$), using Condition 2 as a base level, was greater than the coefficient for the Condition 1 "Near A, Far B, Far C" explanation (4.73, 95% CI $[3.21, 6.26]$). Similarly, these data were only about three times more likely to be observed ($BF_{10} = 2.81$) if the estimated interaction coefficient for the Condition 3 "Near B" explanation (1.20, 95% CI $[-0.30, 2.75]$) was greater than the coefficient for the Condition 3 "Far from A and C" explanation (0.71, 95% CI $[-0.62, 2.02]$).

These mixed results suggest that subjects were taking rational support into account and may have been taking simplicity into account as well, but not in the way prescribed by the decision net model. This led us to consider an alternative model.

The non-probabilistic model. All three models' predictions are based on probabilities assigned to the decision net explanations. But some previous studies have suggested that people do not reason about explanations in a normatively probabilistic way (Douven & Schupbach, 2015; Douven & Mirabile, 2018; Mirabile & Douven, 2020). For instance, in our task, subjects may have combined information about rational support and simplicity by computing their product, as in Equation 3, but without normalizing the probabilities afterward. If so, their resulting judgments would not conform to a true probability distribution.

This approach would lead to some different predictions. For example, consider the explanation "X wanted to sit near A and near B". Because sitting near A and sitting near

[5]In R model syntax, we fit the model: `rating ~ condition * explanation + (1|subject)`.

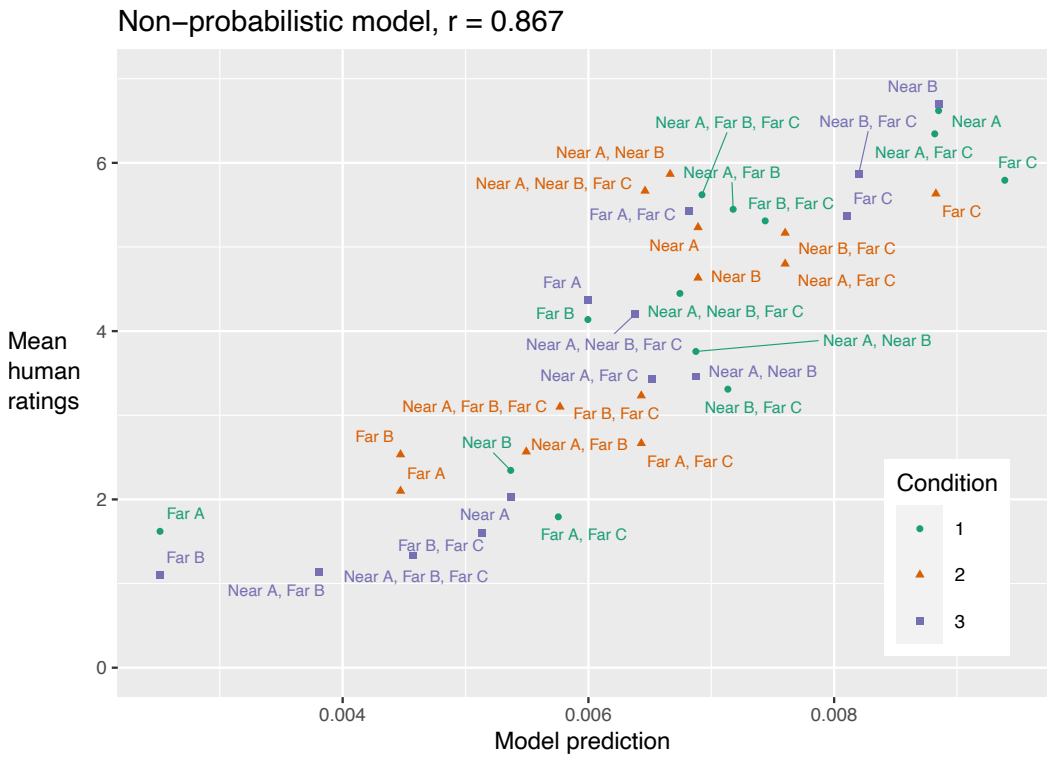

Figure 5. Comparison between mean subject ratings and non-probabilistic model predictions in Experiment 1A.

B are somewhat conflicting goals, each of the seats between A and B achieve this combined goal reasonably well and have similar utility. As a result, the full decision net model assigns approximately equal and relatively low probabilities to X choosing to sit in any of these seats. But if subjects think only about X's seat choice in terms of raw utility and whether the choice accomplished X's goal, X's choice in either Condition 1 (sitting closest to A) or in Condition 2 (sitting midway between A and B) would receive high utility, regardless of the fact that X could have received almost the same utility by making a different choice. This observation is reflected in Figure 4, which shows that subjects rated the "Near A, Near B" explanation much higher than the decision net model in Conditions 1 and 2.

We therefore considered the following alternative model that simply computes the product of rational support and simplicity:

$$v(N|c) = P(c|N) \cdot P(N). \tag{7}$$

Note that here we use the notation $v(N|c)$ because the non-probabilistic model does not compute a true probability. Larger values simply correspond to "better" explanations.

Non-probabilistic model predictions were generated using the same procedure as for the other models. The comparison between best-fitting model predictions and ratings are shown in Figure 5. As the figure shows, the non-probabilistic model predicted subjects'

mean ratings better than any of the other models ($r = 0.87$, best-fitting $k = 0.25$).

Confirming the example above, the figure shows that the non-probabilistic model gives much higher scores to the explanations "X wanted to sit Near A and Near B" in Conditions 1 and 2 relative to other explanations. In Condition 1, the non-probabilistic model also assigns similar scores to the explanations "X wanted to sit near A", "X wanted to sit far from C", and "X wanted to sit near A, and far from C". As Figures 3a and 5 show, subjects also gave these explanations about equal mean ratings. By contrast, the rational support-only model assigned very different probabilities to these three explanations (see Figure 4).

Discussion

The results of Experiment 1A suggest that subjects relied on both simplicity and rational support when making judgments about behavior explanations. Additionally, these principles can be accurately formalized using the decision net framework. Contrary to our predictions, subjects did not combine the principles of simplicity and rational support through probabilistic model selection. However, Experiment 1A was limited by its relatively small sample size and between-subjects design. We therefore conducted a replication with greater statistical power.

Experiment 1B: Replication of Experiment 1A

The method and analysis plan for this experiment was pre-registered. The pre-registration plan may be viewed at `https://aspredicted.org/vw7px.pdf`.

Method

Participants. 85 subjects completed the experiment on the website Prolific. Data from 41 additional subjects were omitted because the subjects failed the attention check. All subjects were paid.

We planned to collect data from at least 85 subjects because a power analysis indicated that 85 subjects would be sufficient to detect, with 95% power, a statistically significant difference in mean ratings assigned to the "Near A" and "Near A, Far from B and C" explanations in Condition 1 that was at least half as large as the effect size we found in Experiment 1A.

Design and Procedure. The experiment was identical to Experiment 1A except that it was within-subjects: each subject saw all three conditions in a random order followed by the attention check. The order of the explanations was randomized for each condition. Like in Experiment 1A, approximately half of subjects saw mirror images of the diagrams in Figure 3 and appropriately modified explanations.

Results

Subject ratings. Figure 6 shows subjects' mean ratings. With some minor exceptions in the rank ordering of the explanations by mean rating, the qualitative results of Experiment 1A replicated.

We repeated all Experiment 1A analyses. The estimated coefficient for the rational support effect was 24.3, 95% CI $[23.2, 25.4]$; the estimated coefficient for the simplicity effect

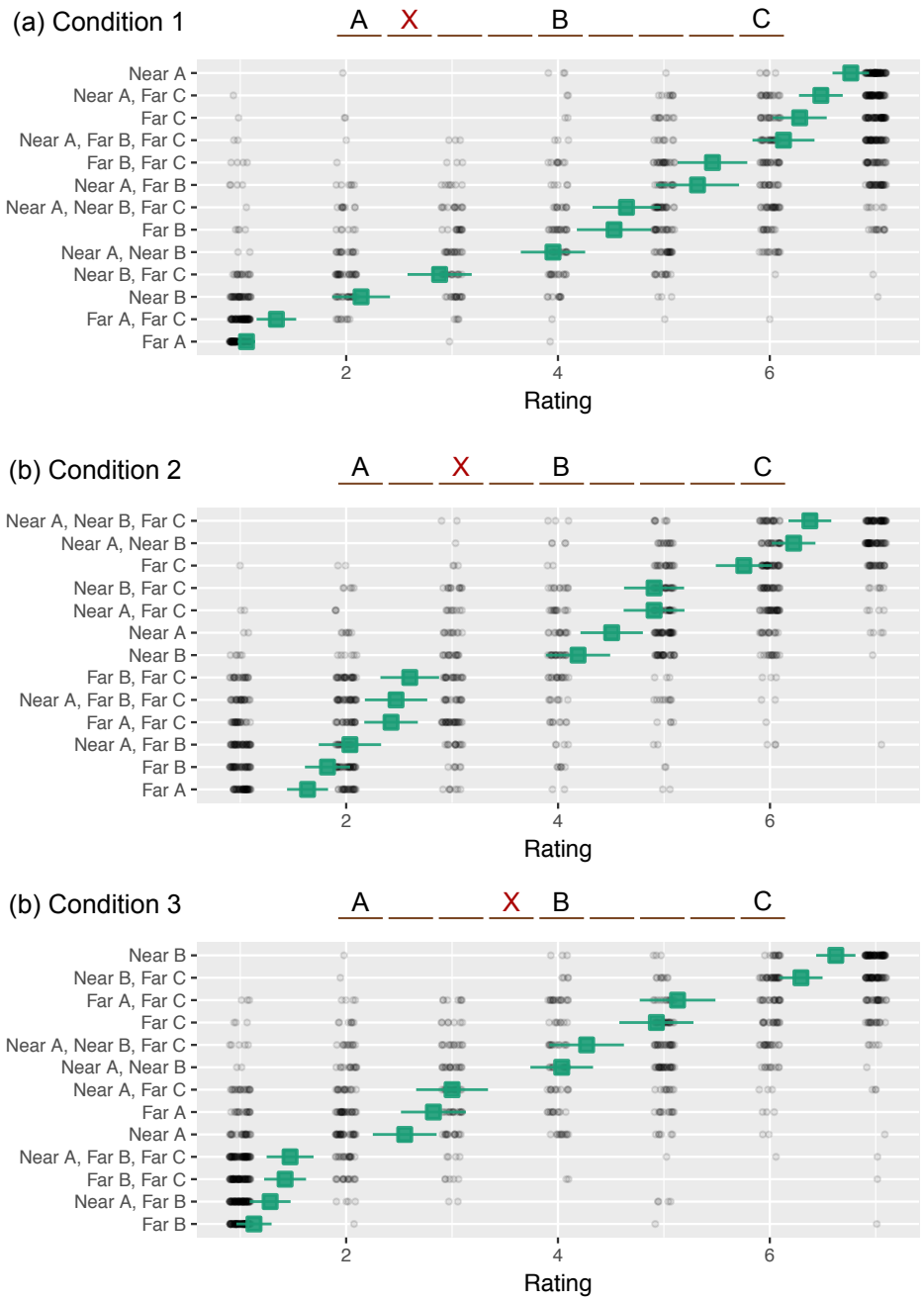

Figure 6. Subjects' mean ratings for how satisfying each explanation was for each condition in Experiment 1B. In each condition, the explanations are ordered by mean rating. The error bars are 95% confidence intervals. The black points indicate individual subjects' responses. Some horizontal jitter has been added to these points to make them more easily visible.

was 33.6, 95% CI [21.9, 45.6]. Both credible intervals did not include 0, replicating the result from Experiment 1A.

In Condition 1, the difference between the mean ratings for the "Near A" and "Near A, Far from B and C" explanations was statistically significant, according to a one-tailed paired-sample t-test, $t(84) = 3.51, p < 0.001$ (difference in means of 0.64, Cohen's $d = 0.56$). In Condition 3, the difference in mean ratings for the "Near B" and "Far from A and C" explanations was also statistically significant, according to a one-tailed paired-sample t-test, $t(84) = 7.73, p < 0.001$ (difference in means of 1.49, Cohen's $d = 1.10$). Both outcomes replicated results from Experiment 1A.

Because Experiment 1B was run within-subjects, we modified our Bayesian ordinal regression to also fit separate slope parameters, as a function of condition, for each subject[6]. Contrary to the results of the t-tests, our data were only slightly more likely to be observed ($BF_{10} = 0.42$) if the estimated interaction coefficient for the Condition 1 "Near A" explanation (8.84, 95% CI [7.33, 10.55]), using Condition 2 as a base level, was greater than the coefficient for the Condition 1 "Near A, Far B, Far C" explanation (9.11, 95% CI [7.73, 10.73]). However, unlike the results of Experiment 1A, these data were about 40 times more likely to be observed ($BF_{10} = 40.84$) if the estimated interaction coefficient for the Condition 3 "Near B" explanation (2.66, 95% CI [1.79, 3.55]) was greater than the coefficient for the Condition 3 "Far from A and C" explanation (1.78, 95% CI [0.96, 2.62]).

Although more than 30% of subjects were excluded from analysis for failing the attention check, the results of all statistical tests were qualitatively the same when all 126 subjects were included in the analyses.

Model predictions. Figure 7 compares subjects' mean ratings with the best-fitting model predictions for all four models. The relative performance of the four models was the same as in Experiment 1A, with the non-probabilistic model performing best of all ($r = 0.85$, best-fitting $k = 0.25$).

Discussion

Experiment 1B largely replicated the results of Experiment 1A with much greater statistical power. We therefore feel confident concluding that our subjects took both simplicity and rational support into account when judging behavior explanations but not in the way predicted by our decision network model.

In both experiments, the non-probabilistic model predicted subjects' judgments better than the decision network model, strongly suggesting that subjects were not making judgments in our task using normative probabilistic inference. If true, this result is somewhat consistent with research on abductive inference—inference to the best explanation. For example, Douven and Mirabile (2018) found that subjects' ratings of the quality of explanations predicted their agreement with the explanations better than their probability ratings that the explanations were true. Other studies (e.g., Mirabile & Douven, 2020) have found similar results, which could support the conclusion that people are not strictly probabilistic or Bayesian when reasoning about explanations (Douven & Schupbach, 2015).

[6] In R model syntax, we fit the model: `rating ~ condition * explanation + (1 + condition|subject)`.

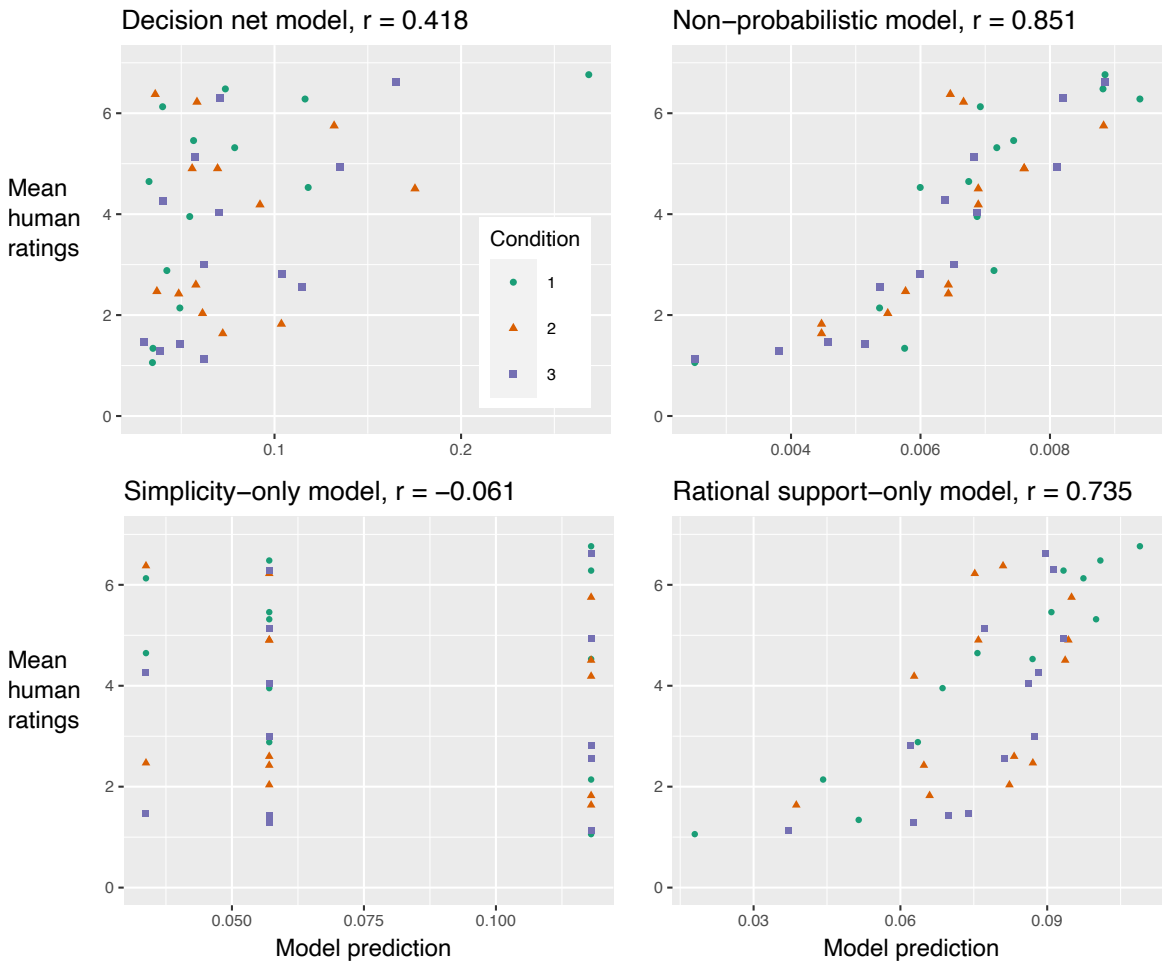

Figure 7. Comparison between mean subject ratings and model predictions in Experiment 1B. Each point corresponds to one explanation from one condition. The x-axes indicate each model's predicted probabilities.

However, another conclusion supported by a study like the one by Douven and Mirabile (2018) is that the "goodness" of an explanation is a concept consisting of multiple aspects, including how satisfying it is and how probable it is. Douven and Mirabile (2018) found that subjects' quality ratings were different from their probability ratings, perhaps because people think of these as independent characteristics. The discrepancy we observed between our decision net model predictions and our subjects' ratings might therefore be due to the fact that we were attempting to predict subjects' *quality* ratings using *probability* predictions. We will revisit this idea in the General Discussion. For now, we turn to a limitation of Experiments 1A and 1B.

Experiment 2: Generating behavior explanations that differ in what someone likes and dislikes

In Experiment 1, we provided subjects with a finite set of explanations to rate and provided all models with the same set of explanations. It is possible, however, that subjects had other explanations in mind that we did not ask them to rate that they would have considered to be even more satisfying. To test this, we conducted a follow-up experiment in which subjects generated their own explanations.

Method

Participants. 45 subjects completed the experiment on Amazon Mechanical Turk. All subjects were paid.

Design and procedure. The design and procedure were identical to Experiment 1A except for two differences. First, subjects saw an example case before the experiment began in which there were only five total seats and three seated people: Person A and Person B were seated in the leftmost and rightmost seats and Person X was seated next to Person A. Subjects were told that X wanted to sit near A and far from B. Second, once the experiment began, rather than rate provided explanations, subjects generated their own explanations.

As in Experiment 1A, subjects saw one of the three cases in Figure 3. They were then asked to provide their best explanation for why person X chose the seat. Subjects were told that X liked some of the people at the meeting, disliked some, and was indifferent toward some. However, no guidance or constraints were placed on subjects' responses. Then then completed the attention check. All subjects passed the attention check and therefore no subjects' data were omitted.

Results and Discussion

Subjects' explanations were coded as one of the 13 explanations in Table 1 or as "other". For example, the response "X doesn't like C" was coded as "Far from C", and the response "X doesn't like anyone" was coded as "other". We (the first and third authors) coded responses independently with 96% agreement (Cohen's $\kappa = 1$ for Condition 1, $\kappa = 0.91$ for Condition 2, $\kappa = 0.90$ for Condition 3, and $\kappa = 0.95$ overall). We disagreed on two responses (one response in Condition 2 and one response in Condition 3) and resolved both disagreements by coding the responses as "other" in order to be as uncharitable to our model as possible. Overall, only 6 of the 45 generated explanations were coded as "other". These results suggest that our list of explanations in Table 1 encompasses most explanations that participants naturally considered.

To further test our models' predictions, we examined whether people were more likely to generate explanations that the models judged to be more probable. The results of this analysis for both the decision net and non-probabilistic models (using best-fitting predictions from Experiment 1B) are shown in Figure 8. The figure shows that, for both models, the top five explanations accounted for more than half of subjects' generated explanations in all conditions. For the non-probabilistic model, the top six explanations accounted for at least 70% of subjects' explanations in all conditions. These results provide additional support

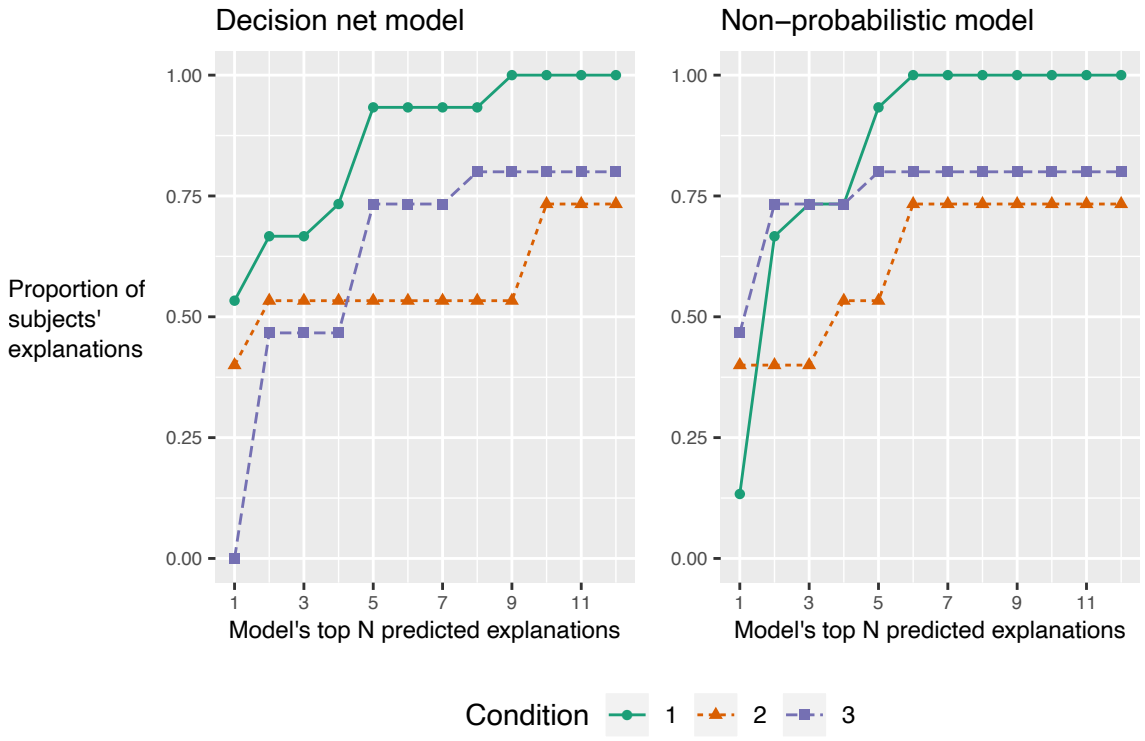

Figure 8. Results from Experiment 2 comparing each model's most probable or most highly rated explanations to the proportion of times subjects' generated that explanation in the experiment.

for our computational approach and reinforce the conclusion that the list of explanations we considered was adequate for this task.

Experiment 3: Rating behavior explanations that differ in what someone knows

Experiments 1 and 2 only examined explanations that differed in what someone likes or dislikes, but our computational framework also makes predictions for explanations that differ in what someone knows. The primary purpose of Experiment 3, therefore, was to test the generality of our computational framework for other types of behavior explanations. Specifically, the task we presented to subjects in Experiment 3 was similar to the one in Experiment 1, but the story and explanations were different.

The method and analysis plan for this experiment was pre-registered. The pre-registration plan may be viewed at `https://aspredicted.org/kw5ab.pdf`.

Didn't know	Clown on A	Clown on A, magician on B, acrobat on C
	Magician on A	Clown on A, acrobat on B, magician on C
	Clown on B	Magician on A, clown on B, acrobat on C
	Magician on B	Acrobat on A, clown on B, magician on C
	Clown on C	Magician on A, acrobat on B, clown on C
	Magician on C	Acrobat on A, magician on B, clown on C

Table 2
Possible explanations used in Experiment 3.

Method

Participants. 100 subjects completed the experiment on Amazon Mechanical Turk. Data from 15 additional subjects were omitted because the subjects failed an attention check described below. All subjects were paid.

Design and Procedure. Subjects first read the following:

> "Four unrelated people, Jacob, David, Alex, and Matt, were all given a coupon to redeem for a free ticket to a show. The show has three stages with different performers. A clown performs on one stage, an acrobat performs on one stage, and a magician performs on one stage. Each person went online and chose one section (of five) to sit in. All stages are at least partially visible from every section, but closer stages are easier to see."

On each trial of the experiment, subjects made judgments about different explanations for why each person chose a ticket for a specific section.

Subjects were randomly assigned to one of two between-subjects conditions. In one condition, subjects were told on each trial that each person "likes clowns and is indifferent toward acrobats and magicians" and in the other condition, subjects were told on each trial that each person "dislikes clowns and is indifferent toward acrobats and magicians". We will therefore refer to these two conditions as the *likes-clowns* and *dislikes-clowns* conditions.

Each subject saw multiple unique trials. In each trial, subjects saw a diagram like the ones in Figure 9. The spaces in the top row labeled A, B, and C, denote the three stages. The spaces in the bottom row denote the five possible seating sections. The section with the X in it denotes where the person chose to sit. Subjects were told: "[Person X] chose the section marked above. Rate how satisfying the following explanations are for why [Person X] chose to sit there." Subjects were presented with the 13 explanations in Table 2, all of which differ in what the person believed (or knew) about where the performers would be at the show. These explanations are expressed in shorthand. For example, "Didn't know" is short for "[Person X] didn't know where any of the performers would be" and "Clown on A, magician on B, acrobat on C" is short for "[Person X] believed that the clown would be on stage A, the magician would be on stage B, and the acrobat would be on stage C". This set of explanations includes every possible belief one could have about the locations of the three performers. Subjects rated each explanation on a scale from 1 ("Very bad explanation") to 7 ("Very good explanation").

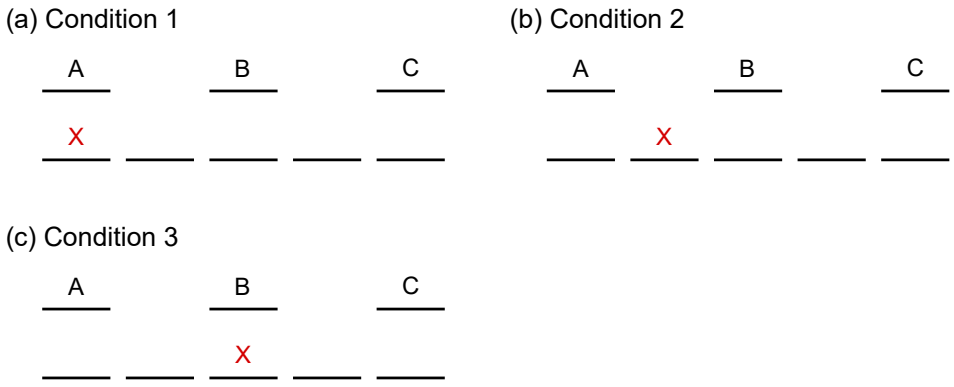

Figure 9. Stimuli used in the three within-subjects conditions in Experiment 3. The first row indicates stages A, B, and C. The second row indicates the five seating sections and the section in which person X chose to sit.

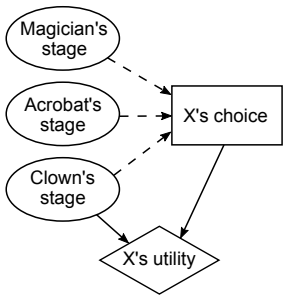

Figure 10. The decision net model used to represent the explanations in Experiment 3. The different explanations in Table 2 differed in the presence or absence of the dashed edges.

Figure 9 shows the three within-subject conditions. Thus, the experiment had a 2 (likes clowns vs. dislikes clowns) × 3 (seating section) mixed design. Like in Experiment 1, approximately half of the trials (randomly) were mirror images of the diagrams in the figure. For example, in Condition 1, the person identified by X would have chosen to sit in the rightmost section rather than the leftmost section.

Attention check. After completing the three trials, subjects were presented with a fourth trial that appeared identical to the previous trials (with a randomly selected diagram) but with the following instructions: "This case is only meant to check if you are reading the instructions carefully. Please ignore the information and leave all of the questions blank." Data from subjects who did not follow the instructions on this trial were excluded from analysis.

Model. Figure 10 shows a template decision net upon which all the explanations in Table 2 are based. These explanations differ in what the person knew or believed about where the performers would be during the show. This is reflected by the fact that the edges leading from the stage locations of the performers to Person X's choice (indicating what X knew before making the choice) are dashed. However, in all cases, subjects were told that

the person choosing a seat either likes or dislikes clowns and is indifferent toward magicians and acrobats. This information is reflected by the presence of an edge leading from the clown's stage location to X's utility and no edges leading from the other performer's stage locations to X's utility.

The remaining modeling details were identical to the decision net model used in Experiment 1. We used Equation 4 to compute the utility that a person would get from sitting in a section. Because the stimuli in Experiment 3 span two dimensions (i.e., there is a row of stages and a row of seating sections), we defined the distance d as the Euclidean distance between the stage of a performer and section in which Person X was seated.

Alternative models. For comparison, we again generated predictions from the simplicity-only model, the rational support-only model, and the non-probabilistic model.

Results

We predicted an interaction between condition and explanation on subjects' ratings, such that subjects would provide different overall ratings depending on whether they were told the person liked or disliked clowns. A 2 (between-subjects condition: likes-clowns vs. dislikes-clowns) $\times$ 2 (explanation) mixed ANOVA found a statistically significant interaction, $F(24, 281) = 3.25$, $p < 0.001$, confirming this prediction.

However, as we noted in our analysis for Experiment 1A, analyzing ordinal data with a metric model like an ANOVA may lead to statistical errors. So, on the recommendation of a reviewer, we deviated from our pre-registered analysis plan and conducted a comparable Bayesian ordinal regression analysis. We fit two Bayesian ordinal regression models, both using subject rating as dependent variables. One model used clown preference (likes vs. dislikes), seating position, and explanation as population-level predictors; the other was identical but omitted clown preference as a predictor. Both models also fit separate intercepts and slopes, as a function of seating position, for each subject[7]. We placed a diffuse and uninformative prior probability distribution of $\mathcal{N}(0, 10)$ on all group level coefficients.

We then compared the two models using leave-one-out cross-validation via the `loo_compare` function of the `brms` R package. The difference in expected log predictive density (-1502.4) was much greater than the estimated standard error of the difference (51.1), suggesting that the model that included clown preference provided a better fit (Vehtari, Gelman, & Gabry, 2017), consistent with the results of the ANOVA test.

To test our prediction that subjects would take both rational support and simplicity into account when making their judgments, we again deviated from our analysis plan and repeated the Bayesian ordinal regression model we ran in Experiments 1A an 1B using subject rating as the dependent variable and rational support-only predictions and simplicity-only model predictions as population-level predictors, plus fitting separate intercepts for each subject. The estimated coefficient for the rational support effect was 26.2, 95% CI $[24.7, 27.7]$. This credible interval did not include 0, providing evidence that subjects used rational support in their judgments. The estimated coefficient for the simplicity effect was -0.5, 95% CI $[-2.5, 1.5]$. This credible interval did not exceed 0, providing evidence that

[7] In R model syntax, we fit the model: `rating` $\sim$ `preference * seating_position * explanation + (1+seating_position|subject)`.

subjects were not using simplicity in their judgments in the way we predicted—with greater simplicity corresponding to higher ratings.

The results of all statistical tests were qualitatively the same when no subjects were excluded from analysis.

The conclusions based on the statistical analyses were reinforced by the modeling results. The mean subject ratings and best-fitting predictions for each of the four models for both the likes-clowns and dislikes-clowns conditions are shown in Figure 11. As in Experiment 1, the simplicity-only model performed very poorly ($r = -0.02$ in the likes-clowns condition, $r = 0.07$ in the dislikes-clowns condition) because this model only takes into account the number of factors referenced in each explanation. Additionally, once again, the rational support-only model outperformed the full decision-net model in both the likes-clowns (rational support-only model $r = 0.92$, best-fitting $k = 1.35$; decision net model $r = 0.78$, best-fitting $k = 2.50$) and dislikes-clowns conditions (rational support-only model $r = 0.86$, best-fitting $k = 0.10$; decision net model $r = 0.66$, best-fitting $k = 0.15$). Unlike in Experiment 1, the rational support-only model also outperformed the non-probabilistic model in both the likes-clowns ($r = 0.69$, best-fitting $k = 1.00$) and the dislikes-clowns ($r = 0.62$, best-fitting $k = 0.15$) conditions.

Discussion

Contrary to our predictions, the statistical and modeling results suggest that subjects incorporated rational support into their judgments but not simplicity. In fact, our results indicated that when simplicity was included in the model, its performance dropped. A follow-up experiment, which we report in the Supplementary Materials, found similar results for a new set of stories, suggesting that the results of Experiment 3 were not due to its particular story.

In Experiment 1, the non-probabilistic model predicted subjects' judgments best, but in the current experiment, the rational support-only model predicted subjects' judgments best. This might imply that subjects in Experiment 3 made judgments more in line with normative probabilistic inference than subjects in Experiment 1. However, this outcome could also be explained by the fact that the rational support-only model is the only model we considered that does not incorporate simplicity into its predictions, and our analyses suggested that subjects did not take simplicity into account when making their judgments.

The difference in results between Experiments 1 and 3 raises the question of why subjects did not consider simplicity when judging behavior explanations that differ in what someone knew or believed. One possible answer relates to a fundamental difference between preferences and knowledge. With preferences, it is possible to like, dislike, or be indifferent to something. With knowledge, it is only possible to know something or not know something[8]. In other words, knowledge statements have an either-or property that preference statements do not. As a result, people may infer what others do and do not know from knowledge statements more easily than from preference statements. For example, if you read a statement saying that Jacob believed that the clown would be on Stage A, you

[8]In rare instances, it is possible for something to be unknowable, which is arguably a third category, but as our account is meant to be an account of folk psychology rather than a fully accurate epistemology, we will ignore this possibility.

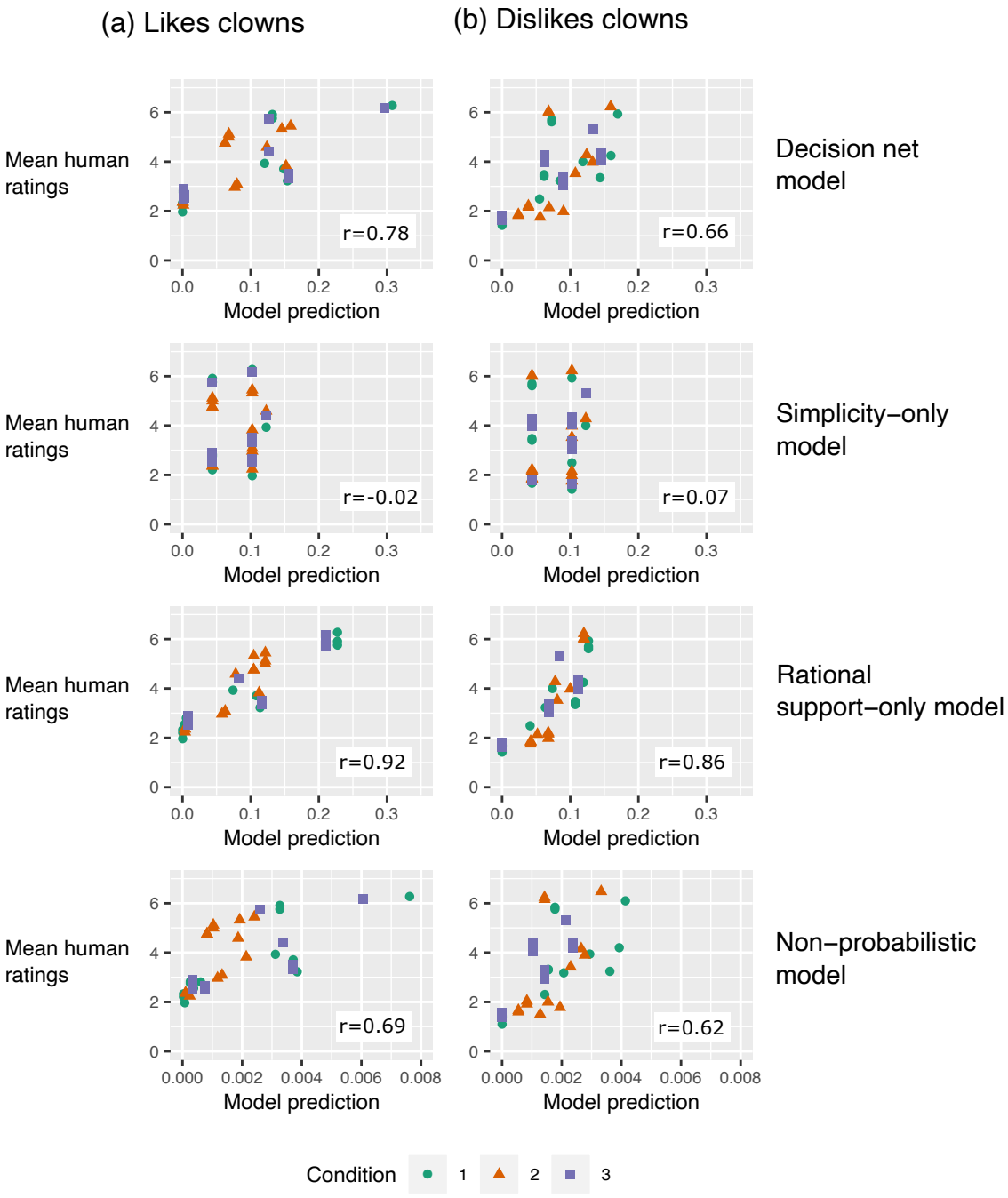

Figure 11. Comparison between mean subject ratings and model predictions for Experiment 3. Each point corresponds to one explanation from one condition. The x-axes indicate each model's predicted probabilities.

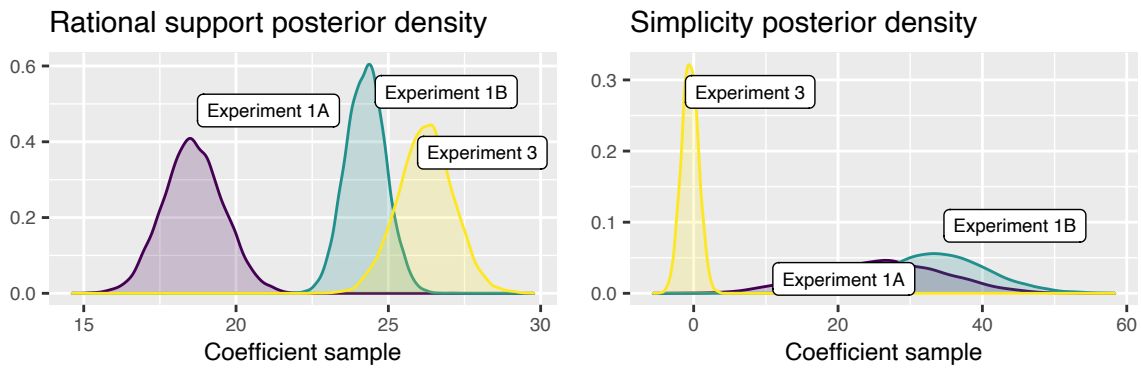

Figure 12. Posterior probability density plots for the two effects of greatest interest from the Bayesian ordinal regression analyses we ran for Experiments 1A, 1B, and 3. The plots summarize the fact that we found support in all experiments for an effect of rational support, but we only found support for an effect of simplicity in Experiments 1A and 1B.

might reasonably infer that Jacob also *did not believe* that the magician would be on Stage B, for example. If he did, the statement would have said so (following Grice's (1975) maxim of quantity for pragmatic communication). By contrast, if you read a statement saying that Lori likes Alice, you are not likely to infer anything about her attitudes toward Bob or Carol.

We call this hypothesis the *knowledge inference hypothesis*. For the explanations in Experiment 3, the knowledge inference hypothesis predicts that subjects would mentally translate the explanations into explanations that refer to all three stages. As a result, every explanation in Table 2 would be equally simple. For explanations that vary only in simplicity, this hypothesis predicts that subjects would rate them all equally, which is what we found in the follow-up experiment reported in the Supplementary Materials.

An alternative possibility is that people assume a lack of knowledge by default, perhaps because ignorance is more common than knowledge. This assumption is similar to the rarity assumption in Oaksford's and Chater's (1994) rational analysis of the Wason selection task, whereby most events in the world are rare (see also Chater & Oaksford, 1999).

General Discussion

We presented a computational framework for understanding people's behavior explanations that formalizes the principles of rational support and simplicity. Our framework constitutes a novel advance toward a formal account of behavior explanation. Across four experiments, we found evidence that people rely on rational support when judging explanations of other people's behavior when those explanations refer to people's preferences and people's beliefs or knowledge. These results are consistent with previous research by Malle (1999, 2004). However, we found inconsistent evidence regarding whether people rely on simplicity when judging explanations of other people's behavior. Specifically, our subjects' took both rational support and simplicity into account when they were reasoning about

preference-based explanations, but only took rational support into account when reasoning about knowledge-based explanations. A summary of the these findings appears in Figure 12.

Additionally, even when subjects did base their judgments on both rational support and simplicity, they did not appear to combine them in a normatively probabilistic way, as predicted by our decision net model. The superior performance of the non-probabilistic model in Experiment 1 adds to an ongoing debate about the role that probabilistic reasoning plays in people's explanatory judgments (Oaksford & Chater, 2020). Balancing rational support (or probability) with simplicity naturally lends itself to a Bayesian formulation, but our Experiment 1 results, as well as the results of several previous studies (Douven & Schupbach, 2015; Douven & Mirabile, 2018; Mirabile & Douven, 2020), suggest that people do not rely only on probability when making judgments about explanation quality.

Regardless of how exactly people make judgments about behavior explanations, our experiments suggest that those judgments rely on the principles of rational support and simplicity. Our decision network account provides a way to formalize those principles that can be applied to any model, as we showed with the non-probabilistic model.

In the following sections, we discuss how our computational framework's emphasis on these two principles can help to account for previous empirical results. We then turn to some limitations of our approach.

Accounting for preferences for conjunctive explanations

In one study, Leddo et al. (1984) found that people preferred conjunctive behavior explanations—explanations that provided multiple reasons—over explantations that provided a single reason. For example, subjects were told that John chose to attend Dartmouth College and rated different explanations for why he chose to go there. One explanation was that "John wanted to attend a prestigious college" and another was that "John wanted to attend a prestigious college and Dartmouth offered a good course of study for John's major". Subjects rated the second explanation as more probable than the first. These results were offered as evidence against the discounting principle of causal attribution (Kelley, 1987) but they could also be offered as evidence against the simplicity principle: why would people rate a more complex explanation as a better explanation than a simpler one?

This example illustrates why our framework highlights the importance of *both* simplicity and rational support. As we explained at the beginning of the paper, our computational framework predicts that one might prefer a less simple explanation if it provides more rational support. Indeed, that might be why subjects showed a preference for the conjunctive explanations in the Leddo et al. (1984) study, especially given that the individual reasons (i.e. wanting to attend a prestigious college) are not sufficient reasons to choose Dartmouth over another college.

Support for this hypothesis comes from the second experiment in their study. In the second experiment, subjects rated explanations for why someone did not take an action. For example, subjects were told that John did not pull over at an Italian restaurant for dinner and rated different explanations for why not. One explanation was that "John was not hungry" and another was that "John was not hungry and John does not like Italian food". In this experiment, the researchers found no preference for conjunctive explanations, perhaps because the conjunctive explanations provided no additional rational support. In

this case, according to our computational framework, simplicity exerted more influence than rational support.

Cases like this one illustrate how our computational framework can clarify empirical results that have previously been the source of confusion in the social psychology literature (McClure, 1998).

"Simplicity" is complex

The results of our experiments suggest that simplicity and the role it plays in people's explanation judgments is hardly simple. Here we consider two aspects of simplicity that appear to influence people's judgments.

At the beginning of this paper, we mentioned the distinction between count simplicity—referring to the total number of causes in an explanation—and root simplicity—referring only to the total number of root causes in an explanation. We ignored this distinction in our experiments but evidence suggests that this distinction influences people's judgments (Pacer & Lombrozo, 2017).

The equivalence of count and root simplicity in our experiments introduces a potential confound in the interpretation of our results. Even if it is true that people do not take simplicity into account when judging explanations that differ in what someone knows, we cannot determine if our results generalize to all factors or only to root factors. But the count-root distinction is not a limitation of our computational account. For example, the simplicity equation, reproduced below, could be modified to sum over only nodes i in the network that represent root factors.

$$S(N) = \frac{1}{\sum_i (x_i \cdot q_i)}$$

In Experiment 1 by Pacer and Lombrozo (2017), this modified approach would explain why subjects favored an explanation for a patient's symptoms that stated that the patient had two different diseases that were both caused by a single underlying disease, rather than two independent diseases. Even though the first explanation invoked more total causes (higher count simplicity) it only invoked a single root cause (lower root simplicity).

Another factor that may affect people's judgments is whether the causes in an explanation are independently sufficient to explain an observation on their own. For example, Lombrozo (2007) found that subjects favored simpler explanations—specifically those with fewer root causes. However, Lombrozo (2007) defined simplicity differently than we did, with more complex explanations consisting of multiple independent causes that were insufficient on their own to explain the observation. In our experiments, more complex explanations included additional causes that were either independently sufficient to explain the behavior (Experiment 1 and Experiment 4) or irrelevant to the behavior (Experiment 3). Therefore, while the causes in all of our experiments were root causes, in some, those causes were independently sufficient causes and in others they were not.

If this distinction between causes that are independently sufficient or not is meaningful, it is one for which our computational framework, in its current formulation, cannot account. But it is still in line with the principles of simplicity and rational support. Identifying sufficient causes of someone's behavior is the same as identifying causes with perfect rational support. Future studies might explore how simplicity and sufficiency interact with

one another and whether the computational principle of rational support also plays a role in people's simplicity judgments of behavior explanations.

Other limitations

Aside from the difficulties with defining simplicity, our framework has some other limitations. One limitation stems from the inherent imprecision of natural language. Namely, there is no algorithm for translating a natural language explanation into a decision net representation. In our experiments, the domains were simple, making these translations straightforward. But that is not the case in other domains. For example, we conducted a small-scale replication of Study 2 from Malle (1999). In the original study, subjects were presented with short descriptions of other people's behaviors. Subjects then rated how intentional each behavior was and generated explanations for the behaviors. For our replication, we selected the ten behaviors with the highest intentionality ratings—these behaviors also resulted in the highest number of reason-based, rather than cause-based, explanations. We presented the set of behaviors to 13 subjects on Amazon Mechanical Turk for payment and asked them to generate an explanation for each behavior. We wished to test whether simpler explanations would be more common, as predicted by our account.

For some behaviors, translating explanations into decision nets to compute their simplicity was straightforward. For example, one behavior was "Anne invited Sue for lunch" and most subjects' explanations were a version of "Anne likes Sue" or "Anne wanted to spend time with Sue". But for other behaviors, translating subjects' explanations was not straightforward. For example, one behavior was "Anne ignored Greg's arguments" and more than one subject provided a version of the explanation "Anne did not want to talk about what Greg is talking about". There are multiple plausible ways to interpret this explanation and translate it into a decision net, and depending on how one translates it will affect how "simple" the decision net explanation is. This example illustrates why our framework will have difficulty in some cases predicting why some explanations are more common than others, even if the account's basic assumption is correct that simpler explanations are more common.

In addition, other studies have found behavior explanation preferences that our computational framework cannot accommodate. Giffin, Wilkenfeld, and Lombrozo (2017) found that when subjects were presented with an unusual behavior, like a woman screaming at and hitting her boss after being approached about a work project, subjects rated explanations of the behavior to be more satisfying when they attributed the behavior to a named condition (e.g., Depathy) than a unnamed condition or to a general tendency to behave in the observed way. In their experiments, the only difference in the explanations was the presence of a name, therefore it is difficult to argue that one explanation was perceived as simpler than the other. Instead, the authors argue that the inclusion of a label in the explanation invites a stronger inference of a causal link between the condition and the behavior. Okten and Moskowitz (2018) found that subjects prefer explanations of behavior that refer to either traits or goals depending on several features of the behavior. Because our framework does not draw a distinction between traits and goals for the purpose of comparing explanations, it cannot explain this result. Our framework should therefore be viewed as a first step toward a complete computational account of how people explain behavior.

Conclusion

Understanding how people explain other people's behavior is not only a question for psychology. A better understanding of this ability also has implications for artificial intelligence (AI) and cognitive science more broadly. The ability to understand and explain other people's behavior is one important aspect of intelligence for which current AI is vastly inferior to humans (Fong, Nourbakhsh, & Dautenhahn, 2003; Breazeal, Buchsbaum, Gray, Gatenby, & Blumberg, 2015; Davis & Marcus, 2015; Lake, Ullman, Tenenbaum, & Gershman, 2017). Cognitive scientist Gary Marcus has proposed that a modern Turing Test (a test of human-level intelligence) should require an AI system to explain agents' behaviors:

> "... allow me to propose a Turing Test for the twenty-first century: build a computer program that can watch any arbitrary TV program or YouTube video and answer questions about its content. Why did Russia invade Crimea? or Why did Walter White consider taking a hit out on Jesse?" (Marcus, 2014)

A complete computational account of how people explain behavior would provide a common language for psychologists and AI researchers and to help close the current social intelligence gap between humans and AI, moving AI closer to being able to pass a test such as this one. Our research suggests that decision nets are a fruitful framework for developing such a computational account.

Acknowledgments

We thank Mike Oaksford and Patricia Mirabile for valuable feedback on the manuscript and Eric Reyes for help with statistical analysis. Part of this work was presented at the 37th Annual Conference of the Cognitive Science Society. The head silhouette images in Figure 1 are from Freepik.com.

References

Anderson, C. A., Krull, D. S., & Weiner, B. (1996). Explanations: Processes and consequences. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles.* New York, NY: Guilford.

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 Behavior*, *1*(0064).

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

Baker, C. L., & Tenenbaum, J. B. (2014). Modeling human plan recognition using Bayesian theory of mind. In G. Sukthankar, R. P. Goldman, C. Geib, D. Pynadath, & H. Bui (Eds.), *Plan, activity and intent recognition: Theory and practice.* Morgan Kaufmann.

Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., & Blumberg, B. (2015). Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life*, *11*, 1–32.

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

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–64.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92–103.

de Campos, L. M. (2006). A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *The Journal of Machine Learning Research*, *7*, 2149–2187.

DeJong, G. (2006). Toward robust real-world inference: A new perspective on explanation-based learning. In *The 17th European Conference on Machine Learning.*

Douven, I., & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1792–1813.

Douven, I., & Schupbach, J. N. (2015). Probabilistic alternatives to Bayesianism: the case of explanationism. *Frontiers in Psychology*, *6*, 459.

Fass, D., & Feldman, J. (2002). Categorization under complexity: A unified MDL account of human learning of regular and irregular categories. In *Advances in Neural Information Processing Systems 15.*

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Flores, M. J., Gámez, J. A., & Moral, S. (2005). Abductive inference in Bayesian networks: Finding a partition of the explanation space. In L. Godo (Ed.), *Symbolic and quantitative approaches to reasoning with uncertainty.* Springer.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*, 143–166.

Giffin, C., Wilkenfeld, D., & Lombrozo, T. (2017). The explanatory effect of a label: Explanations with named categories are more satisfying. *Cognition*, *168*, 357–369.

Gilbert, D. T. (1995). Attribution and interpesonal perception. In A. Tesser (Ed.), *Advanced social psychology.* Boston, MA: McGraw-Hill.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1). Oxford University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics.* New York, NY: Academic Press.

Howard, R. A., & Matheson, J. E. (2005). Influence diagrams. *Decision Analysis*, *2*(3), 127–143.

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

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38.

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

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2). New York, NY: Academic Press.

Keil, F. C., & Wilson, R. A. (2000). Explaining explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition.* MIT Press.

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

Kelley, H. H. (1987). Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 1–26). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology*, *31*, 457–501.

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

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Korman, J., Voiklis, J., & Malle, B. F. (2015). The social life of cognition. *Cognition*, *135*, 30–35.

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Leddo, J., Abelson, R. P., & Gross, P. H. (1984). Conjunctive explanations: When two reasons are better than one. *Journal of Personality and Social Psychology*, *47*(2), 933–943.

Legare, C. H., & Clegg, J. M. (2015). The development of children's causal explanations. In S. Robson & S. Quinn (Eds.), *Routledge international handbook on young children's thinking and understanding.* Routledge.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning.* Oxford University Press.

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

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York, NY: Wiley.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23–48.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* MIT Press.

Malle, B. F., Knobe, J., O'Laughlin, M., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions. *Journal of Personality and Social Psychology*, *79*(3), 309–326.

Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology*, *93*(4), 491–514.

Marcus, G. (2014). What comes after the Turing Test? *The New Yorker*. Retrieved from `http://www.newyorker.com/tech/elements/what-comes-after-the-turing-test`

McClure, J. (1998). Discounting the causes of behavior: Are two reasons better than one? *Journal of Personality and Social Psychology*, *74*(1), 7–20.

Mirabile, P., & Douven, I. (2020). Abductive conditionals as a test case for inferentialism. *Cognition*, *200*.

Morris, M. W., Smith, E. E., & Turner, K. (1998). Parsimony in intuitive explanations for behavior: Reconciling the discounting principle and preference for conjunctive explanations. *Basic and applied social psychology*, *20*(1), 71–85.

Nielsen, U., Pellet, J.-P., & Elisseeff, A. (2008). Explanation trees for causal Bayesian networks. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence.*

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 305–330.

Okten, O., & Moskowitz, G. B. (2018). Goal versus trait explanations: Causal attributions beyond the trait-situation dichotomy. *Journal of Personality and Social Psychology*, *114*(2), 211–229.

O'Laughlin, M. J., & Malle, B. F. (2002). How people explain actions performed by groups and individuals. *Journal of Personality and Social Psychology*, *82*(2), 33–48.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761–1780.

Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. L. (2013). Evaluating computational models of explanation using human judgments. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence.*

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives.* New York, NY: Oxford University Press.

Tauber, S., & Steyvers, M. (2011). Using inverse planning and theory of mind for social goal inference. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2010). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems 22.*

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.

Vulkan, N. (2002). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101–118.

Yuan, C., & Lu, T.-C. (2007). Finding explanations in Bayesian networks. In *The 18th International Workshop on Principles of Diagnosis.*

Appendix

Different utility functions

In the main text, we only considered one decision function (Equation 2) and one utility function (Equation 4) for the model for Experiment 1A. Here we consider two alternative utility function and choice function combinations and show that the overall model predictions are not heavily dependent on this aspect of the model.

First we consider an alternative decision function:

$$P(c = c_i) = \frac{\exp(EU(c_i))}{\sum_j \exp(EU(c_j))}.$$

This is known as the Luce (1959) choice rule and has been commonly used to model choice behavior. Table A1 shows overall model performance for all models using this decision function. As the table shows, performance is nearly identical for all models except the non-probabilistic model, which performs worse, suggesting that the non-probabilistic model is more sensitive to utility assumptions.

Next we consider the possibility of negative utilities when someone dislikes someone else—the model in the text only allowed for positive utilities. We do this by modifying Equation 4 as follows:

$$u_i(s) = \begin{cases} e^{-kd} & \text{if X wants to sit near } i \\ -e^{-kd} & \text{if X wants to sit far from } i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Additionally, because a model with this utility function allows for negative expected utility, we cannot also use the simple softmax decision function, which only guarantees valid probabilities when all expected utilities are positive. Therefore, we also again use the Luce choice rule as a decision function. Table A2 shows overall model performance for all models using negative utilities and the Luce choice rule. None of the models are affected by including negative utilities.

Model	r	Best-fitting k
Decision net	0.432	0.25
Simplicity-only	0.004	–
Rational support-only	0.700	0.70
Non-probabilistic	0.493	0.40

Table A1
Model performance using the Luce choice rule as a decision function.

Model	r	Best-fitting k
Decision net	0.449	0.55
Simplicity-only	0.004	–
Rational support-only	0.692	1.00
Non-probabilistic	0.490	0.45

Table A2
Model performance allowing for negative utilities and using the Luce choice rule as a decision function.