Computational principles underlying people's behavior explanations

AJ Piergiovanni
piergiaj@rose-hulman.edu
Department of Computer Science & Software Engineering
Rose-Hulman Institute of Technology

Alan Jern
jern@rose-hulman.edu
Department of Humanities and Social Sciences
Rose-Hulman Institute of Technology

Abstract

There are often multiple explanations for someone's behavior, but people generally find some behavior explanations more satisfying than others. We hypothesized that people prefer behavior explanations that are simple and rational. We present a computational account of behavior explanation that captures these two principles. Our computational account is based on decision networks. Decision networks allow us to formally capture what it means for an explanation to be simple and rational. We tested our account by asking people to rate how satisfying several behavior explanations were (Experiment 1) or to generate their own explanations (Experiment 2). We found that people's responses were well predicted by our account.

Keywords: behavior explanation; social cognition; decision networks

Every day, we generate explanations for other people's behavior. For example, suppose that you observe Bob as he arrives to a meeting. When he arrives, there are three people already seated at a table, one at the far left of the table, one in the middle, and one at the far right. Bob takes the seat closest to the person on the left. Why did Bob choose to sit there? Perhaps he likes the person on the left, or he dislikes the person on the right, or perhaps he dislikes another person who he knows will later sit on the right. Often, there are many possible explanations for people's behavior. Nevertheless, some behavior explanations are intuitively more satisfying than others. For example, you are likely to find any one of the above explanations more satisfying than *all* of the explanations combined: namely, that he likes the person on the far left, *and* dislikes the person already seated on the far right *and* dislikes the person he knew would also sit on the right.

What makes some behavior explanations more satisfying than others? The example above suggests that simpler explanations are more satisfying. However, a simple explanation must first qualify as a valid explanation in order to be satisfying. In other words, the explanation must provide rational support for the behavior. For example, it would not make sense to explain that Bob sat where he did because he does not like the person he sat next to. Such an "explanation" is not satisfying because a rational actor who dislikes someone will generally avoid that person and therefore Bob's behavior is left unexplained. This example suggests that there are two principles that underlie people's behavior explanations. We will refer to these principles as simplicity and rationality.

The importance of simplicity and rationality in behavior explanation is supported by previous research. When explaining causal events, people prefer explanations that posit fewer causal relationships (Lombrozo, 2007). And when people generate explanations for intentional behaviors, their explanations tend to refer to implicit beliefs and desires that provide rational support for the behaviors (Malle, 1999, 2004). Additionally, research has suggested that, when reasoning about other people's mental states, people expect others to behave generally rationally (e.g., Baker, Saxe, & Tenenbaum, 2009; Ullman et al., 2009; Jern & Kemp, 2011). Dennett (1987) has called this expectation the intentional stance.

Behavior explanation is closely related to what social psychologists call interpersonal attribution, the problem of attributing someone's behavior to either dispositional or situational causes. However, the literature on interpersonal attribution has focused primarily on cognitive processes rather than computational principles (Anderson, Krull, & Weiner, 1996; Gilbert, 1998). Similarly, although previous research on explanation generation suggests that people rely on simplicity and rationality when explaining other people's behavior, these principles have not been formally defined and unified in a computational framework (see, e.g., Keil, 2006). As a result, it is difficult to predict how the two principles will each influence people's judgments when people consider explanations that vary in both simplicity and rationality (see Pacer, Williams, Chen, Lombrozo, & Griffiths, 2013).

In this paper, we present a computational account of behavior explanation that formally characterizes what it means for an explanation to be simple and rational. Our account is based on the graphical modeling framework of decision networks[1] (Howard & Matheson, 2005). Decision networks have been used previously to account for people's inferences about other people's mental states (Jern & Kemp, 2011) but have not been used to account for people's behavior explanations. As we show later, because decision networks can be ordered by network complexity, they can be used to provide a formal definition of simplicity. And because decision networks incorporate a notion of choice utility, they can be used to provide a formal definition of rationality.

We begin by describing the basic properties of decision networks and explain how we use decision networks to define simplicity and rationality. We then test the predictions of our decision network account in two experiments in which people judged or generated explanations of someone else's behavior.

Explaining behavior with decision networks

We will briefly introduce decision nets (short for decision networks) using the example situation described at the beginning of this paper. This situation can be represented by the decision net in Figure 1a, where Bob has been replaced with X. X's seat choice is represented by the rectangular

[1] Decision networks are sometimes called influence diagrams.

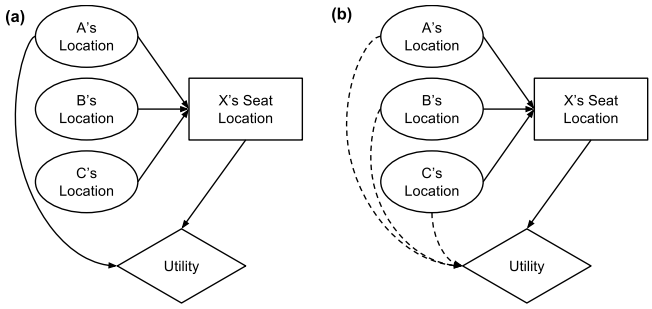


Figure 1: Decision networks. (a) A decision network representing a choice in which only A's location affects X's choice. (b) The set of decision network explanations we considered in our experiments. The networks differed in which of the dashed edges were present.

choice node labeled "X's Seat Location". In this example, X's choice might depend on where Persons A, B, and C are already seated. Their seat locations are represented by the oval nodes. The edges leading from the oval nodes to X's choice node indicate that X knows the values of these variables before choosing where to sit. X's utility is represented by the diamond utility node. The edges leading to the utility node indicate which variables affect X's utility. In Figure 1a, there are edges leading to the utility node from A's location and X's choice. This structure is consistent with X wanting to sit near A and not caring about how close he is to B and C. A fully parameterized decision net would also include a utility function that specifies exactly how X's utility depends on his seat and A's seat, as well as conditional probability tables for any probabilistic variables.

Decision nets assume that choices are taken to increase utility. In Figure 1a, X's utility depends on his choice and A's location. If we suppose that X would like to sit near A, X's utility will be higher for seats that are located closer to A.

The decision net in Figure 1a assumes that X only cares about A's seat location when choosing where to sit. Suppose, however, that you don't know what motivated X's choice, but you observed where he sat and want to explain his choice of seat. This problem is analogous to observing the value of a choice node in a decision net and determining the network structure that best explains the choice. Accordingly, we propose that behavior explanations can be represented by decision nets. We assume that each potential explanation for a behavior corresponds to a fully parameterized decision net in which the value of the choice node has been observed. We now show how, with this assumption, decision nets can capture the principles of simplicity and rationality and can be used to make probabilistic judgments about which potential explanations better explain a given behavior.

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 explanations will have fewer nodes, edges, and possible node values. This intuition can be captured using a definition of simplicity based on minimum description length (MDL; Rissanen, 1978). MDL has been used previously 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 MDL-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 decreases, as the number of edges decreases, and as the number of possible values of each node decreases.

Rationality

Because decision nets assume that choices are taken to increase utility, it is straightforward to capture the principle of rationality. Namely, a decision net explanation provides more rational support for a choice if that choice results in more utility for the actor. A decision net explanation will provide complete rational support for a choice if the choice results in the maximum possible utility for that decision net.

Comparing explanations

We will treat the problem of judging which explanations are better than others as a model selection problem in which the models under consideration are fully parameterized decision nets corresponding to the different explanations. Specifically, we compute the probability that each decision net N is the explanation for choice c:

$$P(N|c) \propto P(c|N) \cdot P(N). \tag{2}$$

In order to compute the likelihood function, $P(c|N)$, we make use of the decision net assumption that actors are likely to make choices that increase their utility. We will consider two ways of instantiating this assumption: by assuming that actors make choices to maximize expected utility, or by making choices probabilistically in proportion to expected utilities. We define the prior probability, $P(N)$, to be proportional to the simplicity of the decision net: $P(N) \propto S(N)$.

Equation 2 shows how decision nets can be used to incorporate formal definitions of simplicity and rationality into computations about which behavior explanations are better than others. Earlier, we suggested that formal definitions of simplicity and rationality allow for predictions to be made about how these two principles will collectively influence people's judgments about behavior explanations. We tested the predictions of our decision net account in an experiment in which people observed someone's choice and judged explanations of the choice that varied in both simplicity and rationality.

One Person	Two People	Three People
Near A	Near A and B	Near A and B, Far from C
Near B	Near A, Far from B	Near A, Far from B and C
Far from A	Near A, Far from C	
Far from B	Near B, Far from C	
Far from C	Far from A and C	
	Far from B and C	

Table 1: The set of possible explanations in Experiment 1.

Experiment 1

In Experiment 1, participants read about a choice someone made and then rated how satisfying several explanations for the choice were. Specifically, participants read about a meeting in which three people, A, B and C, had already arrived and selected seats. Person X arrived last and chose a seat. Participants were told that X likes some people, dislikes some people, and is indifferent toward some people. Accordingly, all of the explanations were expressed as combinations of desires to sit near to, or far from, certain people. Table 1 shows the complete set of explanations shown to participants. For example, the second explanation in the second column of Table 1 identified as "Near A, Far from B" was presented to participants as "X wanted to sit near A and far from B."

Model

All of the explanations in Table 1 can be represented by variations of the decision net in Figure 1b. The differences between the explanations can be captured by differences in which edges lead to the utility node (depicted by dashed lines in the figure). For example, in the decision net corresponding to the "Near A, Far from B" explanation, only the "A's location" and "B's location" nodes would have edges leading to the utility node.

Additionally, differences in explanations with identical network structures, such as the "Near A" and "Far from A" explanations, can be captured by differences in their utility functions. We assumed that the total utility $U(s)$ that X assigned to each seat s depended on the seat's proximity to the people X wanted to sit near to and 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:

$$u_i(s) = \begin{cases} e^{-kd} & \text{if X wants to sit near } i \\ 1 - e^{-kd} & \text{if X wants to sit far from } i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

In this equation, d is the distance, in number of seats, from Person i's seat, and k is a free parameter. Equation 3 has the property that there is a larger difference in utility between the more desirable seats than between the less desirable seats. We then made a standard assumption that utilities are additive. That is, X's total utility $U(s) = \sum_i u_i(s)$. Finally, we assumed that X would choose seats in proportion to utility. That is,

$$P(c = s_j) = \frac{U(s_j)}{\sum_k U(s_k)}. \quad (4)$$

We made this assumption because we hypothesized that, in our low-stakes seating story, participants would not expect people to behave completely rationally. However, we considered an alternative model that did assume that people are completely rational.

Alternative models

We compared our decision net model to several alternative models that were designed to test the importance of our assumptions.

Utility-maximizing model The utility-maximizing model tested whether people only consider an explanation to be satisfying if it provides complete rational support for a choice. This model is identical to our decision net model but assumes that people are completely rational. In other words, this model follows Equation 2 but uses a likelihood function that is equal to 1 if an observed choice results in the maximum utility under a given explanation, and is equal to 0 otherwise.

Simplicity model The simplicity model tested whether the rationality principle is necessary to account for people's judgments. This model is identical to our decision net model but does not take into account how probable an observed choice was under each explanation. In other words, this model follows Equation 2 but sets $P(c|N) = 1$.

Utility-only model The simplicity model tested whether the simplicity principle is necessary to account for people's judgments. This model is identical to our decision net model but does not take into account how simple explanations are. In other words, this model follows Equation 2 but sets $P(N) = 1$.

Method

Participants 125 Amazon Mechanical Turk users completed the experiment. 20 were omitted for failing a manipulation check described below. All were compensated.

Design and Procedure Participants were randomly assigned to one of three conditions. The conditions are depicted in the diagrams above the plots in Figure 2. The diagrams show where Persons A, B, C, and X chose to sit in a row of seats at a meeting.

Participants saw one of these diagrams and were instructed to "Rate how satisfying the following explanations are for why X chose to sit there." They were then shown the 13 explanations from Table 1 and rated them on a scale from 1 ("very bad explanation") to 7 ("very good explanation"). The set of 13 explanations included any explanation that would provide complete rational support for X'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.

The order of the explanations was randomized for each participant. Additionally, half of participants in each condition saw a "mirror image" of the diagrams in Figure 2. For example, in the condition in the center of Figure 2, X would be

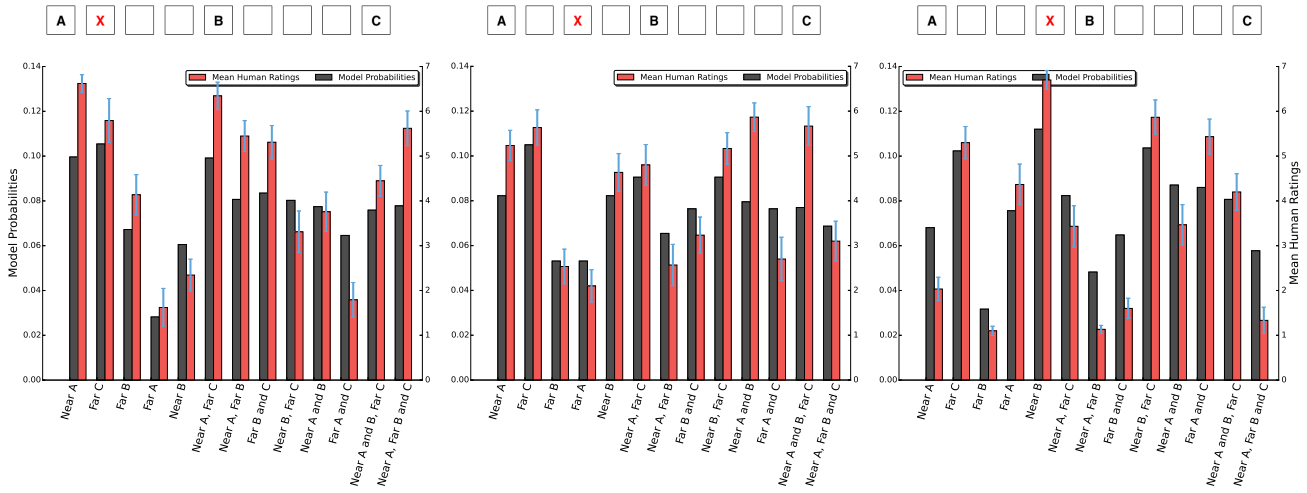


Figure 2: Comparison of decision net model predictions to people's judgments for all explanations in all conditions of Experiment 1.

shown seated closest to C instead of A. For these participants, the set of explanations was adjusted to reflect the different seating location (e.g., "Near A" would be replaced with "Near C"). To ensure that the participants read directions carefully, we included a manipulation check at the end of the experiment. The manipulation check consisted of a second page that appeared identical to the first, but instructed participants to leave all answers blank. 20 participants failed the manipulation check and were therefore omitted from analysis.

Results

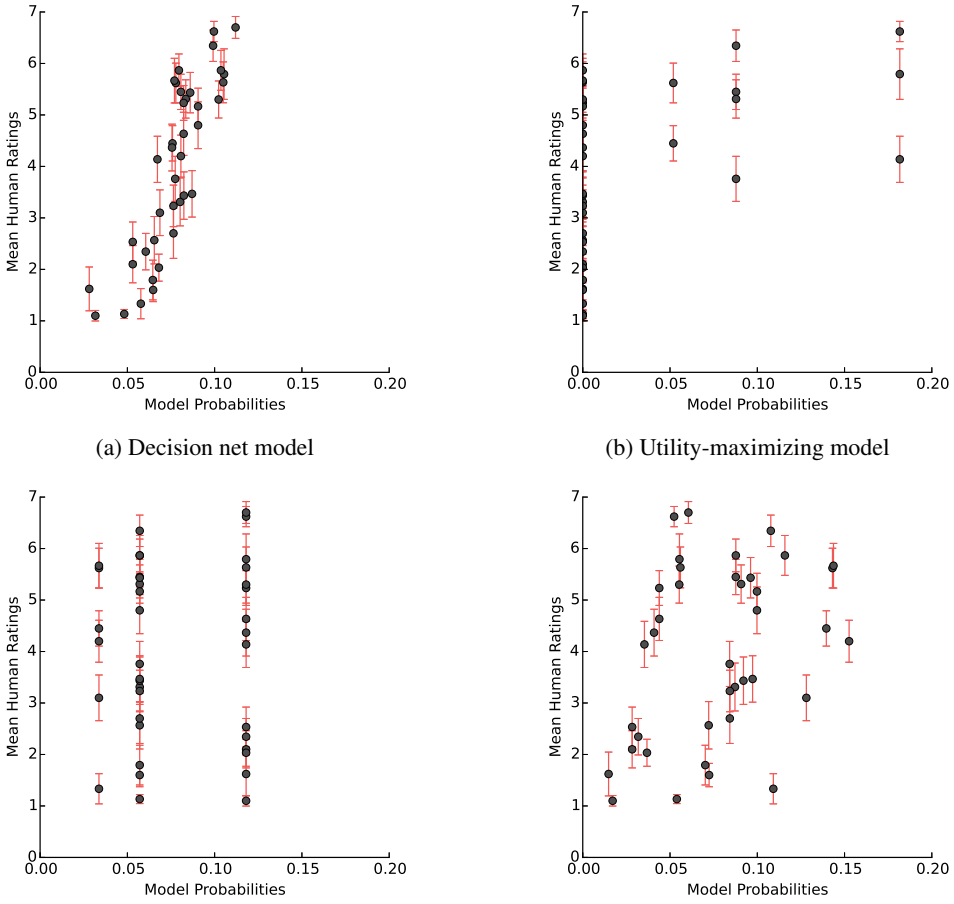
Model predictions were generated by computing $P(N|c)$ according to Equation 2 for each explanation and normalizing the results to sum to 1. When computing $S(N)$ using Equation 1, we assumed that the utility node of each decision net could only take on a finite number of values equal to the number of available seats. Parameter k in Equation 3 was fit to the data for each model (best-fitting k for decision net model: 0.249; utility-maximizing model: 0.245; simplicity model: 0.249, utility-only model: 0.253). Comparisons between model predictions and people's judgments for all 13 explanations across all conditions are shown in Figure 3.

As shown in Figure 3a, our decision net model predicts people's judgments quite well ($r = 0.84$). By contrast, as shown in Figure 3b, the utility-maximizing model performs poorly ($r = 0.53$). The utility-maximizing model assigns a probability of 0 to many explanations that people assigned high ratings to. For example, in the rightmost condition in Figure 2, both participants and our decision net model judged "Far from C" to be one of the top 3 most satisfying explanations. However, the utility-maximizing model assigned "Far from C" a value of 0 because X's seat in that condition is not the optimal seat for being far from C. The poor performance of the utility-maximizing model suggests that people can find behavior explanations satisfying even if they do not provide complete rational support for the behavior.

The predictions of the simplicity and utility-only models are shown in Figures 3c and 3d. Figure 3c shows clearly that people did not base their judgments on simplicity alone ($r = 0.02$). This makes sense when you consider, for example, that the decision nets corresponding to the "Near A" and "Far from A" explanations have identical structure, and are therefore equally simple. However, in most cases, at most one of these two explanations will be reasonable. The utility-only model performs better ($r = 0.34$), but not nearly as well as the decision net model. The poor performance of these two alternative models supports our hypothesis that people rely on both simplicity and rationality when explaining behavior.

Figure 2 compares the decision net model predictions to people's judgments for all 13 explanations in each condition. Overall, the model accounts well for the qualitative patterns in people's judgments. However, there are a few predictions the model gets wrong. For example, in the condition where X sat next to A (left plot), people judged the "Far from A" explanation to be about as satisfying as the "Far from A and C" explanation, while the decision net model predicted that "Far from A and C" is more probable than "Far from A". The model's prediction in this case is a consequence of the fact that it treats utilities as additive. According to the model, under the "Far from A and C" explanation, X's seat provides little utility for being near A, but it provides considerable utility for being far from C. The sum of these two utilities is higher than the seat's utility under the "Far from A" explanation, so the model assigns a higher probability to the former explanation. The fact that people did not do this suggests that they may have considered negative utilities, or that people may not always think of utilities as additive. Perhaps if one part of an explanation is poor, people may judge the whole explanation to be poor.

Although the decision net model accounted well overall for



(a) Decision net model

(b) Utility-maximizing model

(c) Simplicity model

(d) Utility-only model

Figure 3: Comparison of model predictions and human judgments in Experiment 1.

people's judgments about explanations that we provided, it is possible that our participants may have considered additional explanations not in Table 1. Therefore, we conducted a second experiment in which participants generated their own behavior explanations.

Experiment 2

Experiment 2 allowed us to test whether people considered any alternative explanations in Experiment 1 that our model did not account for.

Method

Participants 45 Amazon Mechanical Turk users completed the experiment and were compensated for participation.

Design and Procedure The design and procedure were identical to Experiment 1 except that participants generated their own explanations rather than rate a list of provided explanations. Participants saw one of the three cases in Figure 2 and were asked to provide their best explanation for why person X chose the seat. Participants 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 participants' responses. Participants also completed the manipulation check from Experiment 1. All participants passed the check, so no participants were omitted from analysis.

Results

All generated 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", while the response "X doesn't like anyone" was coded as "other". We (the two authors) coded independently with 96% agreement (Cohen's $\kappa = 0.95$). We disagreed on two responses and resolved the disagreement by coding both responses as "other" in order to be as uncharitable to our model as possible. As shown in Figure 4, the decision net model's top six explanations in each condition accounted for at least 70% of participants' generated explanations. Overall, only 6 of the 45 generated explanations were coded as "other". These results suggest that the our list of explanations in Table 1 encompasses the vast majority of explanations that participants naturally considered.

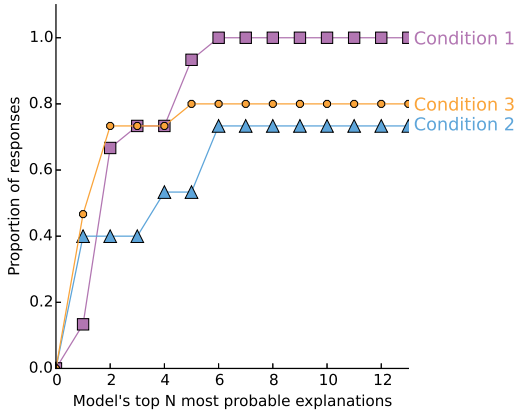


Figure 4: The proportion of generated explanations in Experiment 2 included in the decision net model's top predicted explanations. Conditions 1, 2, and 3 refer to the left, middle, and right conditions in Figure 2.

Conclusion

Our goal was to identify the computational principles that make some behavior explanations more satisfying than others. Overall, our results support the hypothesis that people rely on both simplicity and rationality when judging and generating explanations of other people's behavior and that both of these principles can be formally characterized using decision nets.

Although we considered only a narrow space of possible explanations that differed in utility functions, decision nets can be easily adapted to account for at least one other type of explanation. Recall that some edges in decision nets (such as the edges leading to the rectangular node in Figure 1b) represent what a person knew before making a choice. Consequently, the presence or absence of these edges can be used to represent explanations that differ in what someone did or did not know.

The decision net account in this paper does not provide an account of the cognitive processes involved in explaining behavior, but it may help to motivate future research on cognitive processes. For example, our models assumed that the potential explanations, represented as decision nets, were already constructed and available for comparison. One important question is how those explanations are constructed to begin with. As another example, consider the fact that our definition of simplicity takes into account each decision net's structure. In our experiments, the decision nets had relatively simple structures that could likely be fully stored in working memory. For more complex decision nets, however, people may be limited by working memory capacity. Consequently, people may be unable to fully compare different explanations and will cease to show a simplicity preference.

Overall, our results suggest that decision nets provide a useful formal framework for exploring how people explain behavior. Decision nets provide a formal language for capturing existing ideas from the literature on social cognition and explanation, and present new questions for future research.

Acknowledgments This work was supported by the Rose-Hulman Independent Projects / Research Opportunities Program and ArcelorMittal.

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

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.

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

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.

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

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

Jern, A., & Kemp, C. (2011). Capturing mental state reasoning with influence diagrams. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.*

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254.

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

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*. The MIT Press.

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.

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