Capturing mental state reasoning with influence diagrams

Alan Jern and Charles Kemp
{ajern,ckemp}@cmu.edu
Department of Psychology
Carnegie Mellon University

Abstract

People have a keen ability to reason about others' mental states, which is central for communication and cooperation. A core question for cognitive science is what mental representations support this ability. We offer one proposal based on the framework of *influence diagrams*, an extension of Bayes nets that is suited for representing intentional goal-directed agents. We evaluate this framework in two experiments that require participants to make inferences about what another person knows or values. In both experiments, participants' judgments were better predicted by our influence diagrams account than by several alternative accounts.

Keywords: theory of mind; social cognition; influence diagrams; Bayes nets

People rarely articulate everything they are thinking. Thus, one of the major inductive problems we face is how to infer other people's thoughts from their observable behavior. As a concrete example, suppose your friend asks you to join him in visiting the art museum today, a Monday, which is a day on which you happen to know the museum is closed. You might infer then that your friend does not know the museum is closed on Mondays, or that he does not know that today is Monday, or perhaps neither. You suggest going to the natural history museum instead, which is open but is both more expensive and farther away than the art museum. Your friend declines. Now you might infer that he did not want to spend more money, or that he did not want to travel so far, or perhaps both; it's also possible that your friend simply doesn't like the natural history museum.

People generally find these types of inferences about others natural and exhibit relatively rich intuitive theories of mental states and behavior (D'Andrade, 1987). Mental state reasoning, also called theory of mind, poses some standard questions for cognitive science: Namely, what mental representations support mental state reasoning and what computations are carried out over these representations (Perner, 1991). We propose that these representations are similar to *influence diagrams* (IDs), an extension of Bayes nets that includes a notion of goal-directed action (Howard & Matheson, 2005).

The ID framework provides a graphical language for representing decision problems and an associated formal semantics that supports quantitative predictions. The framework retains all of the strengths of Bayes nets, including the ability to make a distinction between the existence of relationships among variables and the strength of those relationships, to concisely specify a distribution over many variables, and to predict the outcomes of interventions (Sloman, 2005). IDs, however, build on Bayes nets by providing a formal way of making predictions and inferences about intentional behavior.

Computer scientists have previously used IDs to model the behavior of intentional agents, particularly in games (Gal & Pfeffer, 2008; Koller & Milch, 2003), but this research has focused primarily on relatively complex scenarios involving multiple agents, rather than the simple scenarios that have been the subject of most theory of mind research—like the example at the beginning of this paper. And although there are several existing computational models of mental state reasoning (Baker, Saxe, & Tenenbaum, 2009; Oztop, Wolpert, & Kawato, 2005; Schultz, 1988; Wahl & Spada, 2000), the ID framework has received little attention in the psychological literature. We argue that IDs serve as a useful model of the mental representations that support human reasoning.

The rest of the paper is organized as follows. First, we describe the ID framework and discuss some of its strengths for reasoning about other people's mental states and behavior. Then, we apply the framework to a specific task that involves inferring what someone else knows or values and evaluate its ability to predict human performance on the task.

Influence diagrams

We will introduce the ID framework using the following simple scenario. Alice is playing a game in which a two-color die is rolled. Alice chooses a color and receives a reward if her choice matches the color of the die. Thus, there are three variables: the color of the rolled die, Alice's chosen color, and the value of the reward. Two variations of this scenario can be represented by the IDs in Figure 1a and 1b, where the three variables are denoted R, D, and U, respectively.

IDs differ from standard Bayes nets in that they allow for the representation of three semantically distinct types of variables, each of which is shown in the example IDs. First are chance variables, depicted by ovals, which represent probabilistic events like the outcome of the die roll R. Just as in causal Bayes nets, incoming edges to chance nodes represent causal dependencies between events; therefore, we will refer to these edges as causal edges. Second are decision variables, depicted by rectangles, which represent intentional decisions, like Alice's choice D. Incoming edges to decision nodes represent information available when making the decision; we will refer to these edges as knowledge edges. For example, the IDs in Figure 1a and 1b differ in the presence of a knowledge edge from R to D. The ID in 1a represents a situation where Alice knows nothing about the roll before making her choice and the ID in 1b represents a situation where Alice gets to see the rolled color before making her choice. Lastly are utility variables, depicted by diamonds, which represent a decision maker's utility, like Alice's reward U. Incoming

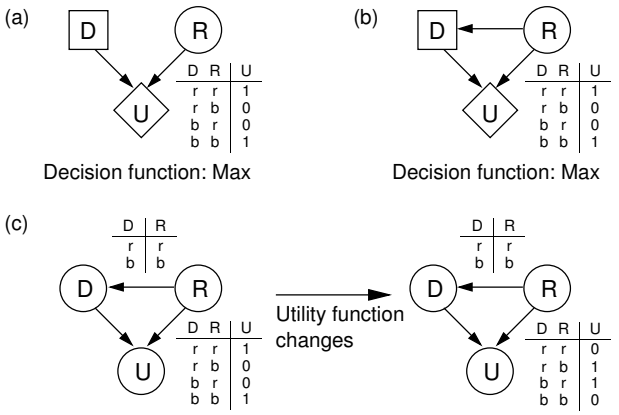


Figure 1: (a) An ID for a single decision with no information. (b) An ID for a decision with complete information. (c) A deterministic Bayes net representation of the ID in panel b. When the utility function changes, the Bayes net incorrectly predicts that Alice will continue to match the rolled color.

edges to utility nodes represent the information that is relevant to the decision maker's state of satisfaction; we will refer to these edges as value edges. In our example IDs, there are value edges from D and R to U representing the fact that Alice's reward depends on the values of both of these variables.

Just as a full Bayes net specification consists a set of conditional probability distributions (CPDs) as well as a graph structure, an ID requires some additional components: a CPD for each chance node, and a utility function that maps the joint values of each utility node's parents to a utility value. In our example, Alice is rewarded only when her chosen color matches the rolled color, as shown in the utility tables in Figures 1a and 1b, where r and b correspond to the two colors red and blue.

Once a utility function is specified, the expected utility EU associated with each possible action d_i can be computed by summing over the unknown variables. For example, when Alice cannot see the outcome c of the roll before making her choice, as in Figure 1a, the expected utility associated with choosing red is $EU(r) = \sum_{c\in\{r,b\}} u(r,c)P(c)$, where $u(\cdot,\cdot)$ is the utility function shown in the table in the figure. However, if Alice is able to see the outcome of the roll before making her choice, as in Figure 1b, there is no uncertainty in the expected utility computation: $EU(r) = u(r,c)$.

The final component of IDs is a decision function σ that specifies the probability of selecting an action d_i for that decision node D. A simple choice of σ is a utility maximizing function, which characterizes the behavior of a rational agent.

$$\sigma(d_i) = \begin{cases} 1, & \text{if } d_i = \arg\max_d EU(d) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

If the die in our example has five red sides and one blue side and Alice maximizes her utility, she will choose red if she cannot see the outcome of the roll, as in Figure 1a. Under some conditions, however, people's behavior is more consistent with probability matching than maximizing (Vulkan, 2002). Thus, another reasonable decision function is a utility matching function.

$$\sigma(d_i) = \frac{EU(d_i)}{\sum_j EU(d_j)} \quad (2)$$

IDs and Bayes nets are closely related, and it is possible to "compile" any ID into an equivalent Bayes net by converting all nodes to chance nodes and choosing CPDs that are consistent with the ID's decision function. For example, the ID in Figure 1b can be compiled into the Bayes net shown to the left of Figure 1c, where the CPD for D is constructed by assuming that Alice acts to maximize her utility. The critical difference between the two representations is that the ID makes the notion of utility maximization explicit, which offers two important advantages. First, the ID representation supports explanations of intentional action (Malle, 1999). If the rolled color is red, then the ID can be used to explain that Alice chooses red in order to maximize her utility. The Bayes net offers no such explanation, and can only indicate that Alice always chooses red when the rolled color is red. Second, the ID representation automatically predicts how Alice's actions will change if the utility function changes. Suppose the game changes and Alice is now rewarded for choosing the *opposite* of the rolled color. After updating the utility function accordingly, the ID representation predicts that Alice will now choose a color different from the rolled color. Figure 1c illustrates, however, that updating the utility node U in the Bayes net leaves the CPD for the decision node unchanged. As a result, the Bayes net incorrectly predicts that Alice will continue to match the rolled color.

Modeling other people's decisions

Although IDs were initially proposed as a way for decision makers to compute optimal decisions under uncertainty, they can also be used to represent other people's decisions. From this perspective, IDs can be used to understand two kinds of mental state inferences: prediction and learning.

Prediction is possible when a person has full information about another person's decision problem, that is, a fully specified ID can be constructed for that person. Predictions can then be made about the utilities that person will assign to possible actions, or the action he or she will take (e.g., by Equation 1). Additionally, because IDs can represent causal relationships using chance nodes and causal edges, it is possible to make predictions about events, just as with standard Bayes nets.

In cases where some details about the decision problem are uncertain or unknown, it may be possible to learn these details by observing the person make some decisions. These situations involve two types of learning problems: structure learning and parameter learning. In terms of IDs, these two problems correspond to learning the graph structure and the ID parameterization, respectively. For example, a person's utility and decision functions may be known but not what information is available when he or she makes a decision. This

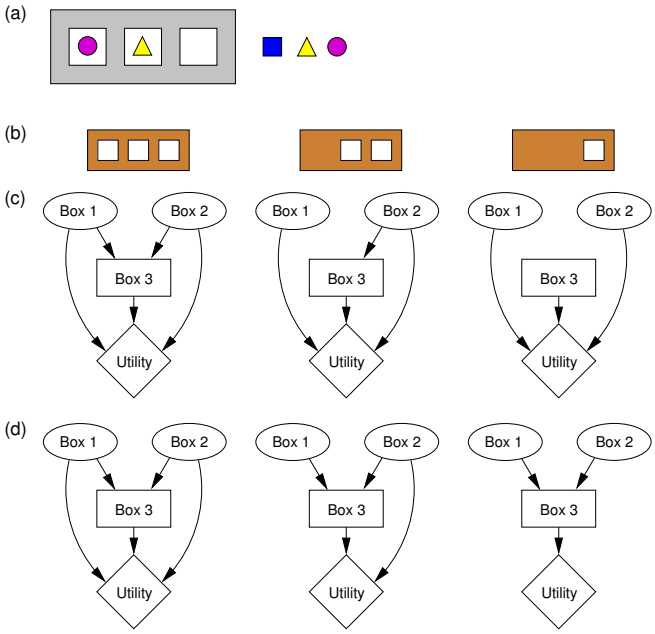


Figure 2: (a) The machine and the three shapes the player may pick from in the shape game. (b) The three different cards in the game. (c) IDs representing the decision problem for each card in Experiment 1. (d) IDs representing the decision problem for each card in Experiment 2.

corresponds to learning what knowledge edges are present in the ID. Similarly, one might learn what a person values (what value edges are present) or what causal dependencies exist (what causal edges are present). Parameter learning applies when an ID graph structure is known, but the precise nature of the relationships between variables is not. This can involve learning the CPDs of chance nodes or the utility functions of utility nodes.

Structure and parameter learning for chance nodes have been previously explored in the context of causal Bayes nets (Griffiths & Tenenbaum, 2005; Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). The remainder of this paper will focus on the less-studied problem of structure learning applied to decision and utility nodes, that is, learning what knowledge and value edges are present in an ID.

Experiment 1

The goal of our first experiment was to examine whether the ID framework can be used to capture how people reason about what other people know. We addressed this question by devising a game called the shape game that allowed us to ask participants about what other players knew during the game based on records of their gameplay.

The shape game

The shape game consists of two components. The first component (Figure 2a) is a machine with three boxes that display shapes. In each round of the game, the machine randomly selects two different shapes from the set of three and displays them in the left two boxes (Boxes 1 and 2). The player then gets to select one of the three shapes to display in the third box (Box 3). The second component (Figure 2b) is a card with holes in it, the "player card", that is placed over the machine at the beginning of the round. There are three different cards: one card covers Boxes 1 and 2 of the machine, one card covers just Box 1, and one card covers no boxes. Thus, depending on the card, the player may be unable to see one or both of the shapes picked by the machine before picking a shape. The goal of the game is to pick a shape that is different from the shapes picked by the machine. Players are awarded 10 points for each pair of mismatching shapes for a maximum of 20 points per round.

In the inference task, a record of 10 rounds from another player is provided, which shows the three shapes from each round but not the card. It is assumed that the same card was used in all 10 rounds. The goal is to infer the card used in the game, based on the player's record of gameplay.

Model

IDs representing the shape game for each card are shown in Figure 2c. In these graphs, the contents of Boxes 1 and 2 are represented by chance nodes, the player's choice for the shape in Box 3 is represented by a decision node, and the awarded points are captured by a utility node. A player's score always depends on the contents of all three boxes, but some cards hide the contents of the machine's boxes before the player makes a decision. Thus, the IDs differ only in the presence of knowledge edges. In other words, inferring the card used involves making an inference about what a player knows, or what knowledge edges are present.

Fully specifying the IDs in Figure 2c requires a decision function that defines a probability distribution over the three options of each decision $d_i \in \{\square, \triangle, \bigcirc\}$. Later we present modeling results based on both the utility maximizing function (Equation 1) and utility matching function (Equation 2). Finally, because rounds are independent, given an ID I_j and a record of n rounds $\mathbf{d} = (d_1, \ldots, d_n)$, $\sigma(\mathbf{d}|I_j) = \prod_i \sigma(d_i|I_j)$.

The inference task can now be framed as a model selection problem where the models are the IDs corresponding to the three cards. We use Bayes' rule to compute the probability of each ID given a set a observed decisions. For an ID I_j, $P(I_j|\mathbf{d}) \propto \sigma(\mathbf{d}|I_j)P(I_j)$. We assume a uniform prior distribution $P(I_j)$, reflecting the fact that all cards are equally probable.

Method

Participants Fifteen Carnegie Mellon undergraduates completed the experiment for course credit.

Design and Materials There are three possible outcomes for each round: all different shapes (outcome D), matching shapes in Boxes 1 and 3 (outcome M1), or matching shapes in Boxes 2 and 3 (outcome M2). It is not possible for the same

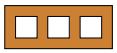
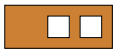
Condition	Gameplay record
	D, D, D, D, D, D, D, D, D, D
	D, D, D, M1, D, M1, M1, M1, D, D
	D, D, D, M1, D, M1, M1, M2, D, D

Table 1: Gameplay records used in the three conditions of Experiments 1 and 2.

shape to be in all three boxes because the machine always picks two different shapes. Participants saw three gameplay records made up of these three outcomes, creating three conditions, shown in Table 1. These conditions were randomly ordered and the specific shapes that appeared in each record were randomly generated for each participant.

These sequences were designed to instill some uncertainty in the earlier rounds about the card being used, but to strongly favor one of the three cards by the final round. For example, in the first sequence consisting entirely of D outcomes, it is possible for a player who cannot see Box 1 or Box 2 to get lucky and choose a mismatching shape every time, but this outcome becomes less likely as the length of the sequence increases. In the third sequence, there is increasingly strong evidence that the player was using the card with two holes until the M2 round, when the one-hole card seems more likely.

The entire experiment was conducted using a graphical interface on a computer. The outcome of each round was shown as a machine like the one in Figure 2a with all three boxes filled with a shape.

Procedure Participants were first familiarized with the shape game by playing six rounds with each of the three cards. Once they indicated that they understood the game, they began the inference task. The sequences of rounds were displayed one at a time with all previous rounds remaining on the screen. After viewing each round, participants were asked to judge how likely it was that the player had been using each of the three cards for the entire sequence of gameplay. They made their judgments for each card on a scale from 1 (very unlikely) to 7 (very likely). They were also asked to give a brief explanation for their judgments.

Results

Model The first model we considered used a utility maximizing decision function (Equation 1). Given the simple nature of the game and participants' own experience with it, we predicted that they would expect other players to generally play optimally. Predictions from this model are shown in the second row of Figure 3a. In the first condition, the model assigns increasing probability to the three-hole card as the number of rounds (all D outcomes) increases. In the second condition, the model rapidly changes its probabilities in favor of the two-hole card after the first M1 outcome. In the third condition, the model raises the probability assigned to the two-hole and three-hole cards after the first M1 outcome, but decreases the probability assigned to the three-hole card until the first M2 outcome is observed, at which point this probability immediately rises to 1.

Human judgments Mean human judgments are shown in the first row of Figure 3a. In order to convert participants' judgments on the 1 to 7 scale to approximate probabilities, the ratings in each round were first decremented by 1 to make 0 the lowest value. Then the ratings were normalized by dividing by their sum to obtain ratings that summed to 1.

In every round of the three conditions, the ordering of participants' ratings is consistent with the model's predictions. Overall, the model captures many of the qualitative trends in the human data, resulting in a high correlation between the human data and the model's predictions ($r = 0.95$). One deviation from the model can be found in the later rounds of the one-hole card condition. Whereas the model predicts certainty in favor of the one-hole card, participants' judgments were less certain and decreased in the final two rounds. This effect, however, appears to have been driven by a subset of participants who took into account the possibility that the solitary M2 round in this condition was a mistake, a possibility that was explicitly noted in several participants' explanations.

Alternative models We compared the human judgments to two alternative models designed to test the importance of the two key components of our model: utility maximization and probabilistic inference. We tested the maximizing assumption by implementing a utility matching model that used a utility matching decision function (Equation 2). This model's predictions are shown in the third row of Figure 3a. Clearly this model offers a poor account of human behavior ($r = 0.61$), especially in the second and third conditions. This suggests that, as predicted, in this simple task, participants assumed that the player they were assessing behaved mostly optimally.

Next, we tested the probabilistic assumption of the first model by comparing it to a purely logical model. The utility maximizing model assigns increasing probability to the three-hole card in the first condition because a long sequence of D outcomes is highly improbable under any other circumstances. This outcome, however, is logically consistent with any one of the three cards. Thus from a logical standpoint, only the M1 and M2 rounds are definitively informative. Predictions based on this approach are shown in the fourth row of Figure 3a. Contrary to the logical model's predictions, however, participants did gradually adjust their ratings on rounds that weren't definitively informative ($r = 0.45$), consistent with the utility maximizing model.

Finally, we examined whether these results could be accounted for by a standard Bayes net structure learning model. Recall that any ID can be compiled into an equivalent Bayes net. Compiling the IDs in Figure 2c into Bayes nets and performing model selection over these networks is one way to implement our ID model. This approach, however, still relies

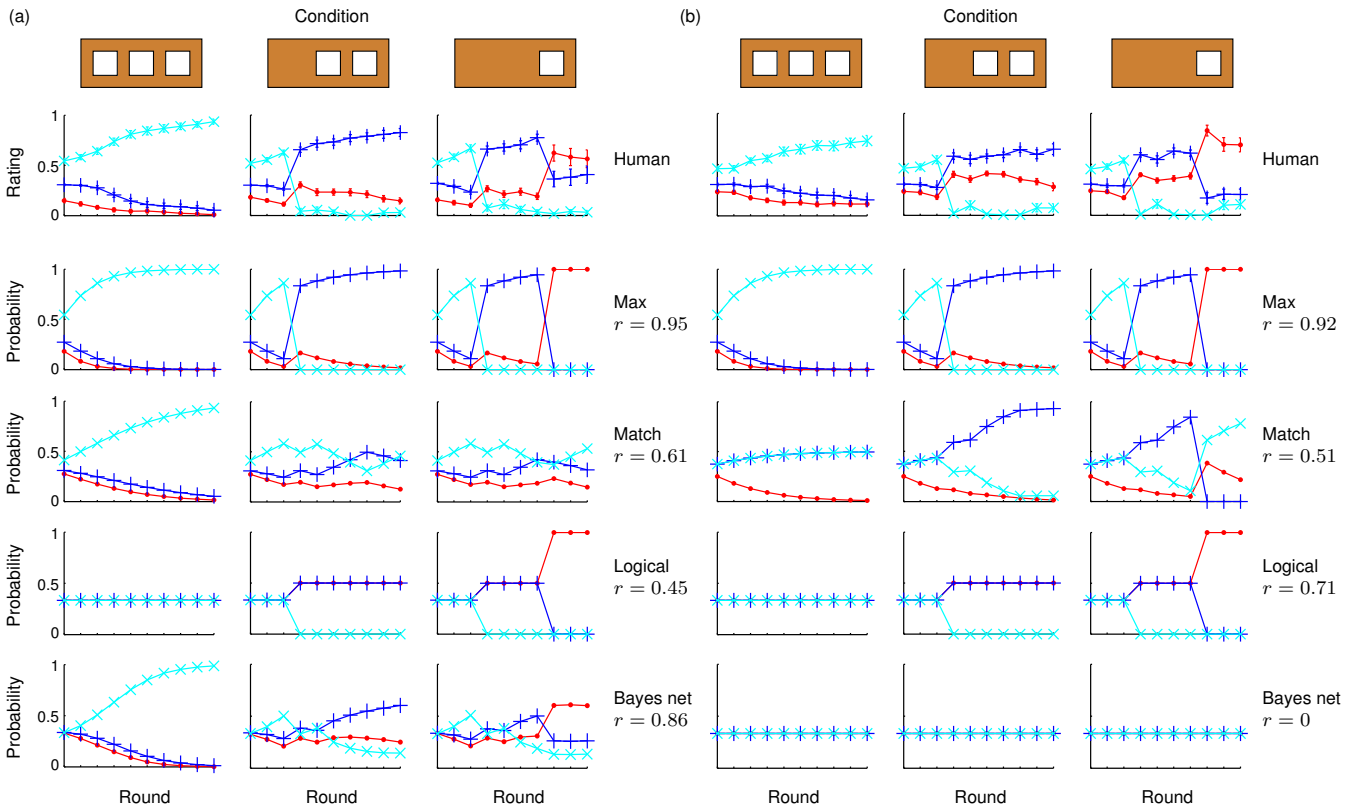


Figure 3: Experiment results and model predictions for (a) Experiment 1 and (b) Experiment 2. In both experiments, the utility maximizing ID model (labeled Max) was the best fitting model we considered. In all plots, the cyan line ($\times$ markers) corresponds to the three-hole card, the blue line ($+$) corresponds to the two-hole card, and the red line ($\cdot$) corresponds to the one-hole card. The error bars in the human plots are standard errors. For the models, r is the correlation coefficient between the model's predictions and the human judgments.

critically on the assumption of utility maximization. We implemented a third alternative model to test whether a Bayes net approach could account for our results without incorporating this assumption. We treated the graphs in Figure 2c as four-node Bayes nets with a known CPD for the utility node and implemented a standard Bayesian structure learning model with uniform Dirichlet priors on the CPDs for the other nodes (Heckerman, Geiger, & Chickering, 1995). The model's predictions are shown in the last row of Figure 3a.

The model performs reasonably well overall ($r = 0.86$) but is inferior to the utility maximizing ID model in two respects. First, after only one round, the Bayes net model assigns equal probability to all three cards, since a single round provides no information about the existence of causal relationships between the boxes. The ID model, however, assumes that the player is attempting to choose a shape for Box 3 that does not match Box 1 or Box 2, and observing a single round where this goal is achieved suggests that the player is able to see both boxes. The second limitation of the Bayes net model is that it fails to predict the dramatic change in participants' ratings after the first M1 round.

Experiment 2

Experiment 1 showed that people's inferences about what another player knew in the shape game were highly consistent with a model selection account using IDs and a maximizing utility function. The purpose of Experiment 2 was to apply this same account to a task involving an inference about what another person values.

Revised shape game

In order to address this question, we made a slight modification to the shape game. In the previous version, the cards were placed over the machine at the beginning of the round. In the current version, the cards—now called "judge cards"—were not placed over the machine until the end of each round. Thus, in the judge card version of the game, players are able to see the shapes in all boxes when making their selections.

The judge card determines how the score for each round is computed: Only the shapes not covered by the card are counted. Thus, when the judge card covers Box 1, the maximum number of points is 10, when the player's shape is different from the shape in Box 2. When the judge card covers Boxes 1 and 2, there are no shapes to mismatch and 10 points are awarded no matter what shape the player picks.

Model

IDs representing the decision problem for each card in the judge card version of the shape game are shown in Figure 2d. The player always gets to see the contents of Boxes 1 and 2, but the awarded points may not depend on the contents of all boxes. Thus, the IDs in Figure 2c differ only the presence of value edges. In other words, inferring the card used involves making an inference about what a player values, or what value edges are present. The remaining details of the model were identical to those in Experiment 1.

Method

All of the participants from Experiment 1 also participated in Experiment 2 (in a random order), with two additional participants whose data from Experiment 1 were lost due to an error (total $N = 17$). Experiment 2 was identical to Experiment 1 except participants made judgments about the judge card version of the game.

Results

The mean human judgments and model predictions are shown in Figure 3b. The utility maximizing and logical models make the same predictions as in Experiment 1. This is because a player who is unable to see the shape in Box 1 is effectively equivalent to a player who does not care about the contents of that box. The prediction that the two experiments produce similar results is largely supported by the human data, which are similar across the two experiments, and utility maximizing model once again performs well ($r = 0.92$). The utility matching model produces different predictions in the two experiments due to the slightly different point assignment policies, but again offers a poor account of the human data ($r = 0.51$). Finally, the Bayes net model is unable to make *any* inferences in the judge card version of the game. This result is a consequence of the fact that observed actions cannot be used to make inferences about a utility function without some assumption about how actions and utility are related (e.g., by a decision function).

Conclusion

The results of our two experiments suggest that people take both decision functions and probabilistic information into account when reasoning about mental states. The different predictions of the utility maximizing and utility matching models supported the idea that people expected others to play nearly optimally, a reasonable expectation in our simple task. However, this utility maximizing assumption alone was not sufficient to capture people's inferences, as indicated by the different predictions of the utility maximizing and logical models.

The influence diagram framework accounted well for both experiments, and performed better than a Bayes net model that did not incorporate the notion of utility maximization. Although Bayes nets share many of the strengths of IDs, they are not naturally suited for reasoning about intentional agents. The influence diagram approach can be viewed as a natural way to supplement Bayes nets with the knowledge that actions are chosen in order to achieve goals. We propose that any successful account of mental state reasoning will need to represent this knowledge in a transparent and explicit way.

References

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

D'Andrade, R. G. (1987). A folk model of the mind. In D. Holland & N. Quinn (Eds.), *Cultural models in language and thought.* Cambridge: Cambridge University Press.

Gal, Y., & Pfeffer, A. (2008). Networks of influence diagrams: a formalism for representing agents' beliefs and decision-making processes. *Journal of Artificial Intelligence Research*, *33*(1), 109-147.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*, 197-243.

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

Koller, D., & Milch, B. (2003). Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, *45*(1), 181-221.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856-876.

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

Oztop, E., Wolpert, D., & Kawato, M. (2005). Mental state inference using visual control parameters. *Cognitive Brain Research*, *22*, 129-151.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Schultz, T. R. (1988). Assessing intention: A computational model. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing theories of mind.* New York: Cambridge University Press.

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

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453-489.

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

Wahl, S., & Spada, H. (2000). Children's reasoning about intentions, beliefs, and behaviour. *Cognitive Science Quarterly*, *1*, 5-34.