Reasoning in teaching and misleading situations

Russell E. Warner, Todd Stoess, Patrick Shafto
Department of Psychological and Brain Sciences, University of Louisville

Abstract

Much of human inference occurs in social situations. While in many cases people cooperate, as in teaching settings, people can misdirect others in order to protect their own interests. Shafto and Goodman (2008) formalized teaching and learning from teachers as Bayesian inference, in which learners use knowledge about the teacher's intent to facilitate inference. This same model provides a basis for exploring reasoning about misleading. We present two new experiments comparing reasoning about teaching and misleading. In both experiments, participants play the role of informant (teacher/misleader) or learner. Our model predicts and our results show that people's behavior differs in teaching and misleading conditions, both when intentions are explicitly known as well as when they are not. Further, the model provides close fits to informants' and learners' behavior.

Introduction

Learning about the world is a daunting task. From the fact that so much of the evidence is underdetermined, to the fact that we have limited time to explore, making inferences about the world is a difficult problem. But what if we are not on our own in this task? Having people around to help us learn about the world might ease some of the difficulty. If knowledgeable informants (mothers, fathers, teachers, friends) helped by choosing the evidence that one saw, learners stand to gain in knowledge about the world–potentially much more rapidly than they could alone.

Indeed, knowledgeable and helpful informants play a central feature of many accounts of cognition and cognitive development (Csibra & Gergely, 2009; Tomasello, Carpenter, Call, Behne, & Moll, 2005; Vygotsky, 1978). For instance, Csibra (2007) claims that children (and not other animals) have an ability to understand intentional teaching as conveying both information about the data, and about the hypothesis that the teaching intends to communicate. This ability is seen as so essential to explaining children's rapid pace of learning, that Csibra and Gergely (2009) suggest that it may in fact be innate.

However, opposite our ability to choose evidence helpfully comes an ability to mislead others with true, but otherwise unhelpful or downright misleading evidence. For this reason, it becomes critical to be able to discern the intentions of individuals sharing information (Sperber et al., 2010).

Consider, for example, the game in Figure 1. In the game, there are concepts (here, boats) of different sizes, and an informant chooses which evidence to supply to the learner. The learner, then must infer the true state of the world, based on the information provided. Clearly, the intention of the informant matters considerably. For instance, in situation A, a teacher would choose to provide the two ends, allowing the learner to infer that the middle must also be a part of the concept. A misleader, on the other hand, would provide either of the other two possibilities, thus leaving the learner uncertain whether A or B/C was true.

Building off of work by Shafto & Goodman (2008), we present a model of teaching/misleading, and inference in each of these situations. We present two experiments testing the predictions of the model, first when the learner knows the informant's intent, and second, when the learner does not know the informant's intent. The results show strong fits to the model's predictions, and show that learners can accurately infer intent based on the evidence alone. We conclude by discussing relationships to other models of inference, and implications cognition.

A model of teaching, misleading, and inference

We formalize reasoning as a problem of probabilistic inference in which learners observe data, d. Given this data, learners update their beliefs about a hypothesis, h, that represents a particular set of concepts. Bayes' rule states that posterior beliefs about hypotheses given data, $P(h|d)$ are proportional to the product of the learner's prior beliefs about the hypothesis, $P(h)$, and the probability of the data given the hypothesis, $P(d|h)$:

$$P_L(h|d) \propto P(d|h)P_L(h), \tag{1}$$

where L indicates learner, and $P(d|h)$ is an appropriate sampling model (e.g. random sampling).

In this paper, we consider data that are sampled by an individual whose intent is to either teach or mislead. That is, we consider informants who choose data intentionally, to either facilitate or impede learning. Our approach builds of that of Shafto & Goodman (2008), who proposed a model of pedagogical data selection. They modeled teaching as choosing data that tend to increase learners' beliefs about the correct hypothesis:

$$P_I(d|h) \propto P_L(h|d)^{\alpha}. \tag{2}$$

where, α was assumed to be greater than 0; thus, the informant chooses data that facilitate inference—they teach. Here, we also consider cases where $\alpha < 0$, the informant chooses data that inhibit learning; they mislead. For our model, in the two conditions, we set these values to -1 and 1.

If the learner is aware that the data are being chosen by another person, then they, in turn, can infer which data a teacher/misleader is likely to choose for any given hypothesis. By substituting their default sampling model with the sampling model used by the informant, Equation 2. Together, Equations 1 and 2 specify a system of equations. We can imagine a process in which the informant and learner each consider each others' inferences, providing a method for solving the system of equations (i.e. fixed point iteration).[1]

An Example of Model Predictions

Consider the concept learning game in Figure 1. In this game, the set of six boats represent the hypothesis space, and each individual boat represents a possible hypothesis. Together this set of hypotheses represent a dimensional concept learning problem (Kemp & Tenenbaum, 2009; Shepard, 1987). In the game, informants choose two windows to display for the learner, and learners infer whether the hidden window is part of the concept (i.e. is hiding a boat).

When the large boat appears, the most helpful move for a teaching informant to make is to reveal the two sides of the ship. This move renders all other ships impossible. As a learner, anytime these data are presented, the inference is simple because all other possibilities have been ruled out. Similarly, for small boats, a teaching informant would choose to send the scenario hidden-water-ship or the reverse, they are eliminating all ambiguities from the scenario. For learners, if they receive this data, inference is again simple. Importantly, learners are more likely to receive this kind of unambiguous situation when dealing with a teacher than when dealing with a misleader.

In contrast, medium ship situations are intrinsically ambiguous, and knowledge of intent (α) plays a more prominent role for learners in these cases. No data pairs provide information that uniquely specifies the hypothesis. In the case that the teacher chooses to expose the windows showing ship sections, what should a learner infer? Although there is no definitive answer to this question, by using their understanding of the teacher's intent learners may nevertheless succeed. In this situation, the only possible hypotheses are the large ship or (one of the) the medium ships. However, the learner knows that the informant is helpful, and if the true hypothesis was the large ship, the teacher would have chosen to present the two edges because those data would be most likely to lead the learner to the correct hypothesis. Therefore, if the learner knows the teacher is helpful, and she is presented with the evidence that one side and the middle are both ships, she should infer that water is behind the hidden window.

[1]The informant must make an assumption about the learner's prior $P_L(h)$. In our experiments both the informant and learner were told that all concepts were equally likely, rendering this inference moot.

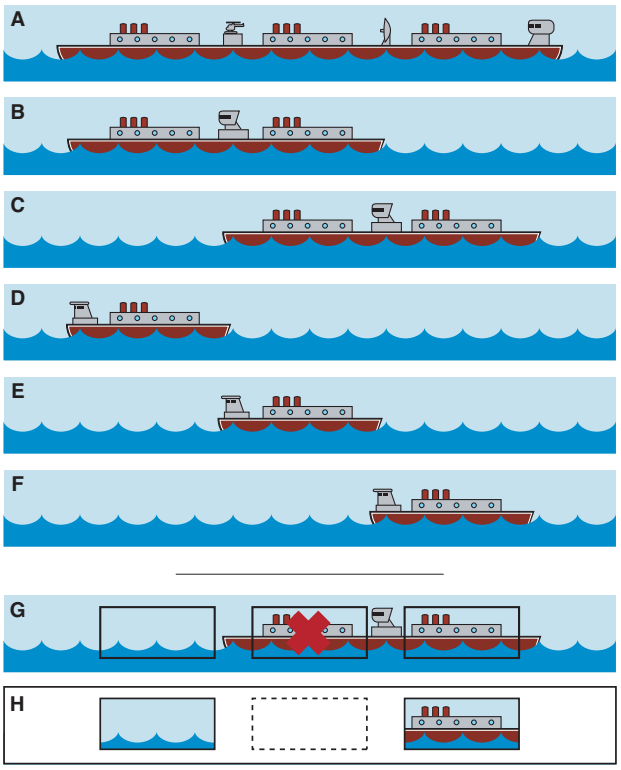

Figure 1: A-F are the six variations on boat positions and types. Informants in both conditions were asked to select a tile to hide from the opposing player (G). The opposing player would then receive an image of the particular scenario with the missing information (H).

Interestingly, by complementary logic, learners in the misleading condition who observe the same information should be less likely to make the same inference. In the case of the large ship, the informant must produce this ambiguous pattern of data in order to mislead, and consequently learners with misleading informants should be more likely to predict that there is a ship hidden.

Experiment 1: Concept Learning with Known Informant Intentions

Experiment 1 tested the model predictions by presenting informants and learners with the aforementioned game. Informants were instructed to either help the learner (teach) or mislead the learner, and learners were told the informant's intent.

Participants. Sixty-four students (32 in each condition) participated in pairs in this experiment in exchange for partial course credit.

Procedure. Participants were asked to play the aforementioned ship game on two computers, obscured from one another by a partition. Individuals were randomly assigned to be a teacher/misleader (hereafter referred to as the informant) and a learner. Before the game, both individuals were shown a piece of paper that served as a key to all possible states of the game (panels A-F on

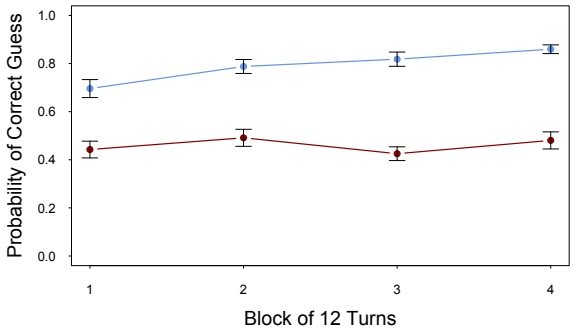

Figure 2: Players in the teaching condition (blue) in Experiment 1 showed an increase in their probability of being correct.

Figure 1) and told explicitly that "only these six states are possible in the game."

In the teaching condition, informants were told that they were to try to choose examples that would "help" their fellow player, and that their score would only increase when the learner guessed correctly. The learner was also explicitly told that their fellow player would try to help them make accurate inferences. Both parties could see one another's scores on the computer screen. In this condition, participants shared a single score, such that when learners guessed correctly, both the informant and learner were rewarded.

In the misleading condition, informants were told that they could only increase their score if the learner was wrong and so they should try to "trick" their opponent. The learner was told of the informant's intentions and could see both player's scores. In this condition, participants had separate scores, such that when learners guessed incorrectly, the informant received points while the learner did not, and vice versa.

In both conditions, the learners provided inferences on a scale that controlled the number of points bet on an inference. Bets could vary from 5 on water, to 0, to 5 on boat. These bets allowed learners to express their confidence in an inference, and for the purpose of analysis we normalized these to a 0 to 1 (water-boat) scale and treated them as probability judgments.

Participants played a total of 48 rounds in which the informant provided data, and the learner made an inference. In each round, informants configured data for the learners by choosing which tile to obscure. Learners then observed this data and made an inference about whether a boat or water was behind the tile. Participants had no communication other than the data. The rounds were divided into four blocks of 12 ships in which each ship appeared twice in a random order.

One pair of participants was removed from the data because due to a computer malfunction, their entire game progress was not recorded.

Results. We conducted preliminary analyses to identify informants and learners who misunderstood the instructions. Specifically, for learners, we identified the six deterministic situations (e.g. SOS, SWO, WOW, etc.), and removed pairs where the learner made more than 2 errors. We also identified informants in the deceptive condition who produced more than 4 of these situations, as this also indicates either non-compliance with or misunderstanding of the instructions. A total of 3 pairs were removed based on these two criteria (two teaching and one misleading condition pairs).

We began by looking at pairs' performance by focusing on the probability of correct inferences to see if learners took advantage of information they had about the intentions of their informant. Learners in the cooperative condition were predicted to have a distinct advantage due to the data they were likely to receive from the informant. Consistent with this prediction, we found that learners teaching condition ($M = 0.87$) performed better than those in the misleading condition, ($M = 0.51, t(27) = 18.34, p < .001$). Given the differences in learners' performance, one may wonder whether there were differences in informants' behavior.

The model predicts that, in the teaching condition, informants should choose helpful moves, such as revealing the two side tiles in the case of the large boat. Conversely, in the misleading condition, we expect that informants would be likely to provide unhelpful information, such as revealing a side and middle tile in the case of the large boat. Figure 3 shows the model predictions and the participants' choices. The model provides close fits to informants' behavior in both conditions ($r = 0.93$ in the teaching condition and $r = 0.98$ in the misleading condition), capturing both qualitative reversals in choices and cases of relative indifference. These correlations are for the full 48 turns, but in this figure we additionally present the first 12 turns to illustrate that qualitatively most of the model's predictions have already begun to take form amongst participants actions. Continuing with the entire 48 turn analysis, in the large boat scenarios, informants in the teaching condition chose to reveal the two sides 98% of the time, $\chi^2(1) = 222.35, p < .001$. This as opposed to informants in the misleading condition who chose to hide a side 99% of the time, $\chi^2(1) = 53.16, p < .001$.

Similarly, the model predicts that learners in the teaching condition should use the information about the teacher's helpfulness to guide their inferences. Learners in the misleading condition should know that the informant will be unhelpful, and use this informant to guide their inferences. Figure 4 shows the model predictions and the results. There were 12 possible scenarios that a learner might encounter. We encoded those scenarios as S for *ship*, W for *water* and O for *occluded*. In some of these scenarios, the inference was straightforward, regardless of condition. In cases such as SOS or WOW, both the model and people agree that S is the most likely inference. Similarly, in cases such as SWO, both the model and people agree that W is the most likely inference. However, in some cases, the data are ambiguous. These ambiguous cases can be grouped into two sets based on the model predictions. Specifically, the model predicts that the SSO and OSS cases are more likely to be W in the teaching than the misleading condition, based on the fact that if the informant was intending to teach about a large ship, they would have chosen SOS.

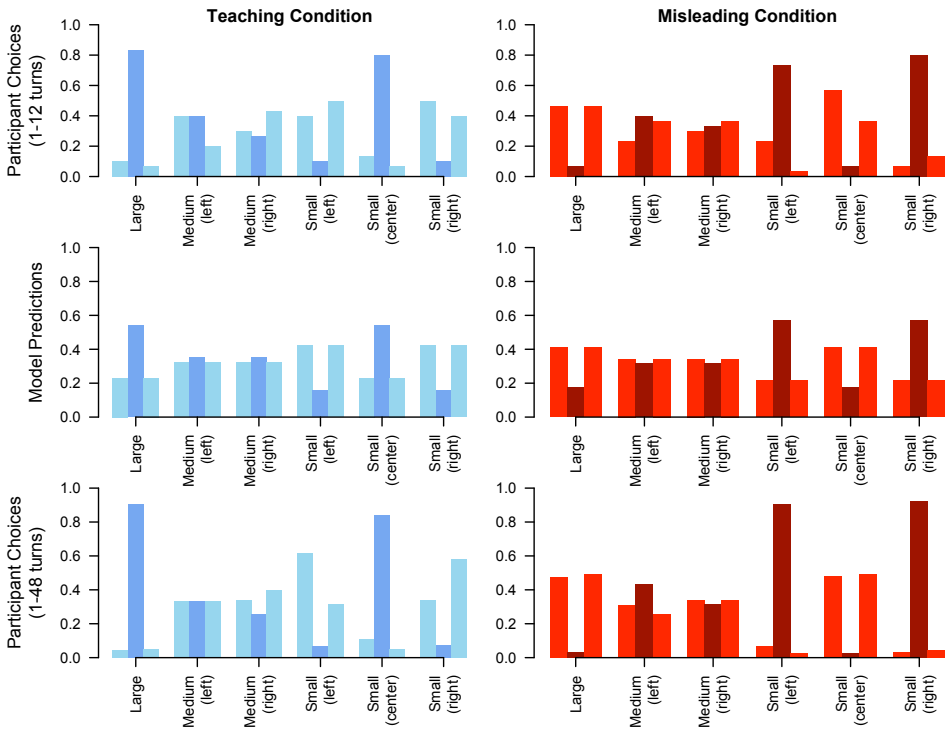

Figure 3: Informants could choose to cover the left, center or right tile. Participants' choices in Experiment 1 are illustrated across the first row, by ship type. The model's predictions of moves are shown across the bottom row.

Conversely, in cases such as SOW, OSW, WSO, WOS the model predicts that the missing case should be more likely to be S for the teaching than the misleading condition.

To test this prediction, for each learner we computed the average probability of guessing Ship or Water for each situation. We tested for the predicted reversal with a 2×2 mixed ANOVA. If people in the teaching condition were more likely to respond Ship to the first cases, and Water to the second cases, and people in the misleading condition responded in the opposite pattern, then we expect a significant interaction. The results confirmed the model predictions, $F(1, 153) = 7.58, p < 0.01$. Overall, the model provided a close fit to people's behavior, with $r(12) = .98$ in the cooperative condition and $r(12) = .93$ in the competitive.

Recall that for these cases (i.e. medium boats), the informants' choices in both conditions appear random (see Figure 3). However, given the high probability of learners guessing correctly in the teaching condition, it seems unlikely that individuals were actually behaving randomly. To investigate this, we coded informants' choices for consistency within the first three and the last three games. Consistency was defined as making the same choice on each of the three games; otherwise, they were coded as inconsistent. In the teaching case, there was a change in the number of consistent choices between the beginning three turns and the last three turns. Participants changed from being inconsistent (3 of 15) to being highly consistent (12 of 15) by their last three turns, $p < .005$ by Fisher's Exact test. In the misleading case, there was no change in consistency between the first three turns (0 of 14) and the last three (1 of 14), $p = 1$. Informants in the teaching condition became consistent as the game progressed, while those in the misleading condition did not.

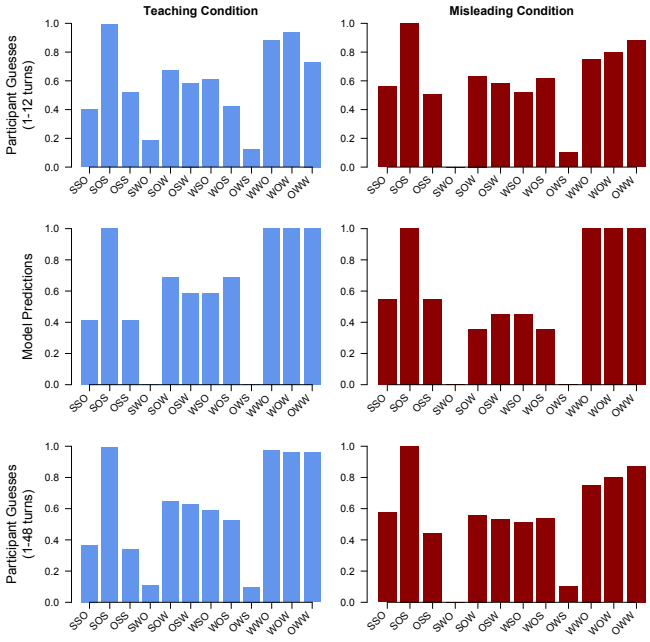

Figure 4: We have labeled possible scenarios as three-letter codes (S = ship, W = water, O = occluded). For instance, if the learner saw SSO, they saw two ship sections on the left and a the third tile was hidden. The bars are the probability that learners in Experiment 1 (or model) would guess that a ship section was present.

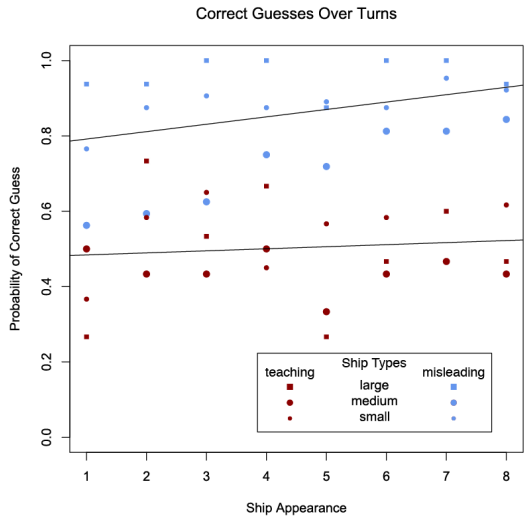

Figure 5: The probability of making a correct guess, per ship type appearance, over the course of a game in Experiment 1.

Because informants exhibit greater consistency over time in the teaching condition, but not in the misleading condition, we expect in increases in accuracy for the teaching but not the misleading condition. Focusing on the intrinsically uncertain cases, the two medium cases (see Figure 5), the probability of guess correct increases, $b = 0.07, t(58) = 2.53, p < .05$. In contrast, for the competitive condition, the probability of guessing correctly did not increase over the course of games $b = 0.02, t(54) = .61, p = .54$. Overall then, learners in the teaching condition were able to utilize their knowledge of their informant's intentions in order to perform better and improve their performance.

Experiment 2: Concept Learning with Unknown Informant Intentions

Having established with Experiment 1 that learners are able to take advantage of teaching conditions when intentions were explicitly known, we turn now to the case where the informant's intentions are not known.

Method

Participants. Ninety-two students (25 in the teaching condition, 21 in the misleading condition) participated in pairs in this experiment in exchange for partial course credit.

Procedure. Our second experiment followed the same structure of the game that participants played in our first experiment with the critical difference being that learners were not told what their partners' intentions were. The informant was told that they were to help or hinder the learner and were shown both her own score and the learner's score. In addition, the learner was specifically told that she could make no assumptions about her partner's intentions, and they were not allowed to observe the informant's score. When participants were finished, learners were asked to rate their informant on a scale from -10 ("extremely deceptive") to 10 ("extremely helpful"). All other details were identical to Experiment 1.

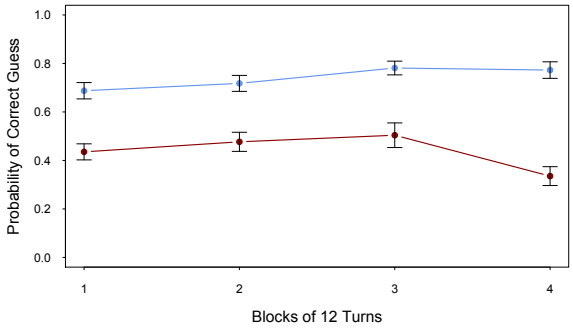

Figure 6: Players in the teaching condition (blue) in Experiment 2 showed a greater probability of being correct. This effect increased with turns.

1.

One pair of participants was removed from the data because the informant explicitly stated after the experiment that she had, "changed her mind about helping her partner" in the midst of the game.

Results. As in Experiment 1, we conducted preliminary analyses to identify informants and learners who misunderstood the instructions. Unlike for Experiment 1, the criteria lead to a number of exclusions (15 pairs). Specifically, 13 learners and 5 informants misunderstood the instructions (in 3 cases these overlapped). After these exclusions, 16 pairs remained in the teaching condition and 15 pairs were in the misleading condition. It appears that uncertainty about the informant's intent lead learners to have difficulty with the instructions. In what follows, we focus on the 31 pairs who passed the manipulation check.

Learners in the teaching condition were more accurate (M=.80) than those in the misleading condition (M=.46); $t(29) = 8.55, p < 0.001$. Learners in the teaching condition rated their informants as more helpful ($M = 3.94$) than those in the misleading condition ($M = -2.47, t(29) = 3.18, p < 0.01$), suggesting that a priori knowledge of intent is not critical to accurate inference.

This difference could simply be a product of receiving better data. If this was the sole explanation, we would would expect constant performance over blocks. Figure 6 shows the probability of correct inferences over the course of the four blocks. Learners in the teaching condition improved over blocks, $b = 0.07, t(62) = 2.79, p < 0.01$, while learners in the misleading condition did not $b = -0.03, t(58) = -1.11, p = 0.27$, suggesting that perhaps learners use inferences about helpfulness to facilitate inference with experience.

To investigate this further, we consider the two kinds of ambiguous data–SSO and SOW–for which the model predicts different inferences in the two conditions. Recall, the model predicts that for cases such as SSO, learners in the teaching condition should infer W because if this was the large ship, then the informant would have chosen SOS instead (and learners in the misleading condition should be more likely to guess S). Similarly, for SOW, learners in the teaching condition should be more likely to guess S than learners in the misleading con-

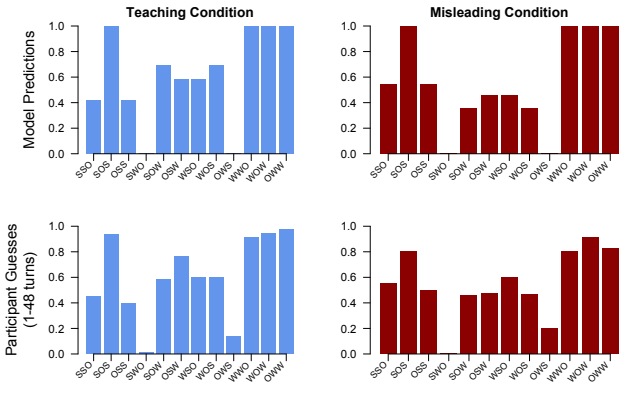

Figure 7: Among the twelve scenarios that players saw in the second experiment—where intent was unknown—the model accurately predicts learner behavior.

dition. To test this prediction, for each learner, we computed the average probability of guessing Ship or Water for each of these kinds of ambiguous situations. We tested for the predicted reversal with a 2×2 mixed ANOVA. If people in the teaching condition were more likely to respond Ship to the first cases, and Water to the second cases, and people in the misleading condition responded in the opposite pattern, then we expect a significant interaction. The results confirmed the model predictions, $F(1, 166) = 12.45, p < 0.001$. Overall, the model predicts people's behavior well, with $r(14) = .97$ in the cooperative condition and $r(13) = .95$ in the competitive. This suggests that learners, with experience, can infer an informant's intent and use it to guide inferences in ambiguous situations.

General Discussion

How do informants' intentions affect how we make inferences about data? We presented a computational model of informants who teach and mislead, and formalized inference in each of these contexts. We presented two experiments, investigating reasoning when the informant's intent is known, and when their intent is unknown. The first experiment showed that the model predicted both informants' choices of data, and learners' inferences, including qualitatively different inferences based on identical ambiguous data. The second experiment showed that learners could infer an informant's intent, and capitalize on this knowledge to support stronger inferences. Taken together, these results provide support for the model of teaching and misleading, and suggest that learners can infer intent based on experience.

Theories of cognition and cognitive development have focused on the importance of pedagogical reasoning in explaining children's ability to learn rapidly from limited data. Our results show that intent can be inferred, and used to guide inferences, based on limited data. This suggests that dedicated pedagogical reasoning mechanisms may not have to be innate, but could potentially be inferred from the input.

Of course, ambiguous intent is hardly the only problem faced when learning from others. Learners must also infer what cues are associated with the intent to teach as opposed to mere intent (or the intent to mislead) (Csibra & Gergely, 2009; Goodman, Baker, & Tenenbaum, 2010). Learners should also be interested in whether individuals are knowledgeable or not (Koening & Harris, 2005). While these inferences are beyond what has been proposed here, this work presents a step in the direction of a richer, more complete understanding of intuitive psychological theories and the role they play in learning.

Our approach represents a step toward integrating formal approaches to understanding learning with formal approaches to games. Here we have focused on showing that people's inferences in communicative situations can be explained by the former approach, and in future work it will be important to explore common and differing predictions made by each framework, as a means toward a more complete understanding of the role of social inference on learning.

One of the great mysteries of human cognition is how we learn so much, so fast, and manage to transmit it effectively across generations. The capacity to share information, do so with limited evidence that can generate powerful inference quickly, and protect ourselves from misinformation goes some of the way toward explaining how it is that we have overcome the massive learning problem that the natural world represents.

References

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Goodman, N., Baker, C., & Tenenbaum, J. (2010). Cause and intent: Social reasoning in causal learning. , 2759–2764.

Kemp, C., & Tenenbaum, J. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Koening, M., & Harris, P. (2005). The role of social cognition in early trust. *Philosophical Topics*, *9*(10), 457–459.

Shafto, & Goodman. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 1632–1637). Austin, Texas: Cognitive Science Society, Inc.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Sperber, D., Clèment, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (2010). Epistemic vigilance. *Mind Language*, *25*, 359–393.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691.

Vygotsky, L. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.