Distinguishing first-line defaults and second-line conceptualization in reasoning about humans, robots, and computers

Daniel T. Levin[a*]
[a]Vanderbilt University, Dept. Psychology and Human Development, Nashville, TN 37203, USA, Daniel.t.levin@vanderbilt.edu

Megan M. Saylor[b]
[b]Vanderbilt University, Dept. Psychology and Human Development, Nashville, TN 37203, USA, M.saylor@vanderbilt.edu

Simon D. Lynn[c]
[c]Vanderbilt University, Dept. Psychology and Human Development, Nashville, TN 37203, USA, Simon.d.lynn@vanderbilt.edu

[*]Corresponding author
Daniel T. Levin
Tel.: +1 615 322-1518
Fax: +1 615 343-9494
Vanderbilt University
Department of Psychology and Human Development
230 Appleton Place #552
Nashville, TN 37203-5701
E-mail: Daniel.t.levin@vanderbilt.edu

Abstract

In previous research, we demonstrated that people distinguish between human and nonhuman intelligence by assuming that humans are more likely to engage in intentional goal-directed behaviors than computers or robots. In the present study, we tested whether participants who respond relatively quickly when making predictions about an entity are more or less likely to distinguish between human and nonhuman agents on the dimension of intentionality. Participants responded to a series of five scenarios in which they chose between intentional and nonintentional actions for a human, a computer, and a robot. Results indicated that participants who chose quickly were more likely to distinguish human and nonhuman agents than participants who deliberated more over their responses. We suggest that the short-response time participants were employing a first-line default to distinguish between human intentionality and more mechanical nonhuman behavior, and that the slower, more deliberative participants engaged in deeper second-line reasoning that led them to change their predictions for the behavior of a human agent.

Keywords: Human-robot interaction, Theory of mind.

1. Introduction

A central requirement for successful human-robot interaction is to understand the concepts that people use when thinking about intelligent artificial agents. Within the general framework of people's understanding of technology, these concepts are particularly important for HRI. Because robots are, in many cases, embodied interactive partners with their human users, they do more than simply respond to specific commands; they interact with people, learn from them, and make intelligent decisions, and it appears as though people sometimes prefer to interact with robots that exhibit characteristically human social cues (Bruce et al., 2002). Accordingly, it is important to create robots that behave in a manner consistent with user expectations. Lee et al. (2005) have explored specific knowledge that people attribute to robots and we have explored what people think more generally about the capabilities inherent to a range of intelligent artificial agents. In previous work, we have established that people strongly distinguish goal-directed human thought from more rote computerized thought, and that they default to the assumption that robots are similar to computers. We established this by asking participants to make predictions about the behavior of a human, a robot, and a computer in a range of scenarios in which these agents interacted with a set of objects (Levin et al., 2008). In the present study we expand on this finding to explore how people use a range of reasoning strategies to draw conclusions about the difference between artificial and human agents. In particular, we created an on-line version of our behavioral prediction scenarios, and tested whether participants who make their predictions rapidly distinguish between natural and artificial agents more or less than participants who make their decisions more slowly. We find that more rapid decision making is associated with a stronger dissociation between human and nonhuman thinking.

1.1. Humans, Robots, and Computers – do they think alike, or differently?

Perhaps the most basic question one might ask regarding people's concepts about natural and artificial agents is whether people believe that different agents "think" in fundamentally different ways. It is possible, for example, that people simply generalize their naive psychology to apparently intelligent machines. One compelling aspect of this possibility is that it allows people to benefit from a well-established "Theory of Mind" which they use to understand human behavior in terms of the beliefs, desires, and goals that drive it (Gopnick and Wellman, 1992). This generalization might, in many ways, be relatively effective for understanding the representations that guide the behavior of artificial agents, and some existing data at least indirectly imply that people do this. For example, experiments by Nass and Moon (2000) demonstrate that people apply a range of social norms to computers. On the other hand, research exploring early-developing foundations of TOM in infants and children demonstrates that they distinguish between goal-directed human action from more mechanical actions that lack these goals (Woodward, 1998).

In this context, we have been exploring adults' beliefs about how different kinds of agents think. Our basic approach has been to ask participants to make specific behavioral predictions about different entities that rely on deeper concepts related to TOM. At the most general level, these are concepts about intentionality. Intentionality is defined in two basic ways, and we draw on both. First, intentionality reflects a level of analysis for human behavior (Dennett, 1991). In theory, one could try to understand why people do the things they do by analyzing the statistical regularities of stimuli preceding specific actions, and the consequences that follow them. An alternative would be to invoke mediating representations of beliefs, desires and goals as a more efficient way of understanding and predicting behavior. Second, intentionality is defined as a specific property of mental representations that affords a close connection between mental symbols and their referents – this reflects the idea that humans have a semantic system that allows them to truly know what a symbol such as a word refers to in the real world, while, according to many philosophers, computers do not (Searle, 1984). Thus, an intentional theory of mind is a collection of specific and general strategies for understanding human behavior using an underlying intentional framework. In the remainder of this paper, we will refer to predictions of human-like actions as being intentional predictions.

In our prediction scenarios, participants are shown a simple scene depicting a set of objects and/or simple events. In some cases, participants are told that some agent has acted upon some of the objects, and then are asked what the agent will do next. For example, one of the scenarios is based on the above-mentioned developmental research demonstrating that infants understand goal-directed action (Woodward, 1998). In this scenario, illustrated in Figure 1, participants first see a pair of objects, a duck and a truck, resting on a simple grid. They are told that an agent has acted twice (by reaching in the case of the human and the robot) upon the duck in location A-1. Then, participants are shown a new image with the locations of the duck and the truck swapped, and are asked whether the agent will act again upon the duck (now in the new location), or upon the truck (now in the location formerly occupied by the duck). If participants believe that

they entity is intentional, and has goals like a human, then they would believe that the first actions were goal-directed reaches to the Duck, and predict that the entity would again reach to the duck in the new location. On the other hand, if the entity had no goals and was instead engaging in more rote behavior, they should predict that it will simply return to the old location despite the fact that it contains a new object.

Other scenarios assess the second aspect of intentionality: the direct, effective connection between mental representation and objects in the world. One of these is founded on the assumption that there is a close link between intentionality and the understanding that words generally refer to categories of objects organized by function, rather than by surface perceptual features (Bloom, 1997). So, participants were shown a set of objects that might be categorized in one way based on meaning/function, and in another based on perceptual features (one of the objects was a candy bar that looked like office supplies, and the other was an office object that looked like a piece of candy). Several experiments have demonstrated that when these scenarios were presented to participants, they made considerably more intentional predictions for human agents than for computers (Levin et al., In review; Levin et al., 2008; Levin et al., 2006). This basic contrast was validated by participants' ratings of computers' ability to infer the goals of human action, and their overall intelligence. In a series of multiple regressions, participants who believed that computers can infer goals showed less of a human-computer intentionality difference than participants who did not believe that computers can infer goals, and this effect was independent of overall ratings of computer intelligence (Levin et al., In review). A key question is how participants construe robots, which share features of both humans and computers. In a number of experiments we have observed that participants initially do not distinguish computers and robots on the dimension of intentionality, even when the robots are given proper names, and shown engaging in human-like locomotion. However, when participants were induced to pay close attention to repeated episodes of a robot looking at one of a pair of objects, they did begin to make more intentional predictions for the robot than the computer (Levin et al., In review).

1.2. The transition model of reasoning about agency

Results such as those described above, combined with previous research make clear that reasoning about agency is not a simple matter of categorizing an entity. For example, we have completed developmental experiments demonstrating that as children become more sophisticated in understanding the difference between living and nonliving things, they borrow and combine features of the two kinds when answering questions about robots (Saylor et al., In press). In addition, previous research has explored participants' understanding of a computer vision system that was described in nonintentional terms, or anthropomorphized by giving it a name, and describing it in terms of goals. In this case, predictions about the anthropomorphized system's capabilities were more human-like, but only when participants could not answer questions based on simple deductions, and were instead forced to reply on deeper concepts (Levin, In review). Combined, all of these findings suggest that a complete understanding of reasoning about agency will require an appeal to a range of reasoning

processes, and specification of the situations in which these processes are likely to occur.

In response to this need, we are developing the transition model of agency-reasoning. This model hypothesizes two modes of reasoning, first-line defaults and second-line conceptualization, and it puts special emphasis on specifying the circumstances under which people transition between modes. The idea of two modes of reasoning is a generalization of models of TOM which specify an initial, automatic agency-detection/belief inference stage of processing, and a later, controlled conceptual stage that does the more difficult job of tracking situations where belief and the world do not align (Apperly et al., 2006; Baron-Cohen, 1995; Leslie et al., 2004). The transition model is similar, but it does not assume a direct link between specific styles of processing and the two stages. Instead, the model distinguishes between one collection of processes that people engage in when they first encounter an agent, and reason about it using existing, unmodified concepts, and another collection that reflects more controlled reasoning, problem solving, and that appeals to deeper concepts about living and nonliving things. Our ultimate goal with this model is not only to specify first-line and second-line processes and knowledge, but also to specify a) the circumstances that induce people to switch from the former to the latter, and b) the consequences of this switch both for deeper concepts and first-line defaults. Thus, the transition model seeks to combine a description of people's understandings about agency with a principled way of explaining how these conceptions might change in response to experience.

Applying the transition model to our finding that people differentiate human and nonhuman thinking, we can ask whether this distinction reflects first-line defaults or second-line conceptualization. On the one hand, it is possible that participants do not have much of a default notion about the intentionality of different systems, and only come to this conclusion with considerable second-line deliberation. On the other hand, if there is continuity between early-developing TOM and adult reasoning about agents, then one would expect that the distinction would reflect an ingrained first-line concept, especially if it is employed commonly. A similar question can be asked of participants' initial failure to distinguish between the robot and the computer (in the absence of the object-attention manipulation): is it the result of a first-line default that can easily be overridden if participants give the system a second thought?

As an initial test of these hypotheses, the experiment we describe here attempts to characterize the predictions people make when engaging in different modes of reasoning by comparing the behavioral predictions people make when they respond quickly to a scenario vs. more slowly. This simple distinction is inspired by variety of research paradigms that rely on the assumption that the time participants take to make a decision can serve as a reliable marker that differentiates the kind of reasoning they have used. For example, research on eyewitness lineup identifications shows witnesses who have responded quickly have replied upon a more direct perceptual recognition, whereas slower identifications reflect a deeper (and generally less reliable) inter-stimulus comparison process (Dunning and Perretta, 2002). Using the same logic, we switched over to computer presentation of our scenarios to allow us to measure the time participants take to make their responses. If the distinction participants make between

intentional human behavior and nonintentional computer behavior reflects first-line defaults, then it may be possible to observe the effect dissipate among participants who deliberate over their decision. A secondary goal of this study was to validate a more dynamic mode of scenario presentation, so instead of telling participants about the initial actions taken by the agents, and instead of showing different categorizations, we showed participants sets of objects along with appearing and disappearing arrows reflecting the actions of each agent in real time. If this method proves successful in replicating our basic effects, it will allow us not only to assess the time participants take to make predictions, but also to manipulate the time they have available to them.

2. Experiment one

2.1. Method

2.1.1. Participants

A total of 30 participants (7 male, three unknown gender) completed experiment one. Their mean age was 34. Participants were students in General Psychology classes at Nashville Community College.

2.1.2. Materials

Participants responded to a series of five agency questions. All five asked participants to imagine three different agents' responses in a specific scenario. Each scenario showed a set of objects that the entities acted upon. The precise nature of the actions (e.g. whether they were looks, points, or some other action) were left ambiguous, and were demonstrated by dynamically appearing then disappearing arrows pointing to the acted-upon object. The action arrows were followed by a pair of choice arrows which indicated two possible actions that might follow the initial actions depicted by the action arrows. The choice arrows were visible until the participant responded. Three of the scenarios were very similar to those used previously (Levin et al., In review). In the first of these, the Duck-Truck scenario, the scene contained a duck and a truck on a simple grid of lettered rows and numbered columns, with the duck in location A-1 and the truck in location C-3. The first two action-arrows indicated successive actions oriented toward the duck. Then, the duck and truck suddenly switched positions, and two arrows appeared, indicating two possible final actions, one to the duck and one to the truck. As reviewed in the introduction, the duck-oriented action would be considered more intentional. In the Pen-Row scenario, a row of seven small objects was depicted on a table. The first, third, fifth, and sixth were all writing utensils. The others were not. The first three action-arrows pointed to pens in the first and third positions, and a pencil in the fifth position. The two choice arrows pointed to a marker in location 6, and a screwdriver in location 7. A choice of the first choice arrow, pointing to the marker in location 6, would be intentional, whereas choosing the screwdriver in location 7 would be a continuation of the spatial response pattern and would be scored as nonintentional. In the third scenario, the Feature-Category scenario, six objects were shown on a desktop: three were candy and three were office supplies. A dark rectangular candy bar was intended to

look like two of the office objects (a dark rectangular PDA, and a dark rectangular eraser), and a small colorful push pin was intended to look like two of the candies (a gummi bear and a hard candy). The first two action-arrows pointed at the eraser, and the PDA, and the choice arrows pointed to the candy bar and the push pin. The intentional response would be to predict an action to the same-category (but perceptually different) push-pin.

Figure 1. Illustration of the Duck-Truck scenario. The green arrow represents one of the first two "actions" taken upon the duck by the entity.

The final two scenarios were new, but generally tested similar ideas. The Coin scenario showed, from right to left, a penny, a nickel, a dime, and a quarter. The action arrows pointed to the penny, then to the nickel, and the choice arrows pointed to the quarter and the dime. Choosing the quarter would be a nonintentional interpretation that the entity was acting on objects of increasing size, and the intentional response would be to pick the action on the dime, suggesting actions based on knowledge about value. Finally, the Candy-Paper scenario depicted a set of six objects. Three of them rested upon a different piece of paper. The first two action arrows pointed to pieces of candy on paper, and the choice arrows pointed to a video converter plug on paper, and a roll of Smarties, resting directly on the desktop. The intentional response would be to predict an action oriented to the Smarties, which represented a category-consistent action, ignoring the fact that the previous objects had rested on paper.

2.1.3. Procedure

Scenarios were presented to the participants on 13-inch color LCD laptop displays. Before responding to the five scenarios, participants read on-screen instructions that first briefly introduced the entities: "a person named John, a robot called ASIMO, and a computer system called Yd3", and emphasized that "we are asking for your intuitions, and that there are no right or wrong answers - just respond based on your judgment about what each thing will do". These instructions also noted that ASIMO could physically grab things with his arm, and that Yd3 "has been loaded into a system that can physically lift objects at different locations using a mechanical vacuum device". A second screen of instructions explained to participants that they would see objects, and arrows depicting actions each entity had taken upon those objects. They were told that their job would be to view 2-3 actions, and then to choose from a pair of alternatives, which action would follow.

After reading the instructions, participants completed the five scenarios in one of two orders (forward or reversed). Within each scenario, participants made three responses, one for each entity. The order of entities was counterbalanced across participants. For each scenario, participants saw a picture of the first entity and then hit a key to indicate their readiness to continue. Next, they saw 2-3 action-arrows followed by choice-arrows, made their response, and the cycle repeated for the other two entities. Once participants responded to all five scenarios, they gave demographic information, completed a brief survey assessing beliefs about technology, and completed the Need For

Cognition scale.

2.2. Results and Discussion

First, we tested whether each participant's median response time (RT) predicted differences in responding between humans and computers, and between robots and computers. RTs, along with participant age (which we have previously observed to be correlated with intentionality responses (Levin et al. In review)) and Need for Cognition, were therefore entered as predictors into two multiple regressions, one for the difference in intentional responding between the human and the computer, and one between the human and the robot. The regression predicting the human-computer difference demonstrated that increases in RT were significantly predictive of smaller human-computer differences in intentionality (Beta=-.507, $p<.025$), while age (Beta=-.064) and Need for Cognition (Beta=-.201) were not. The regression predicting the difference between human and robot intentionality was similar: increased RTs predicted a smaller difference between humans and robots in intentionality (Beta=-.504, $p<.001$). Increased age was also predictive of a lessened difference (Beta=-.572, $p<.05$), while Need for Cognition scores were not (Beta=.056).

To clarify the pattern of results, participants were divided into two groups by RT (the fastest 15 vs. the slowest 15; see Figure 2). This makes clear that participants who responded relatively quickly indicated that the human agent would respond significantly more intentionally (72%) than the computer (43%; $t(14)=2.662$, $p=.019$) or the robot (36%; $t(14)=5.077$, $p<.001$). The difference between the computer and the robot was nonsignificant. Those who responded more slowly showed no significant differences in predicted intentionality between the agents, although there was a nonsignificant trend for the computer to be judged as more intentional (57%) than the human (43%; $t(14)=1.749$,$p=.102$), and the robot (43%; $t(14)=1.852$, $p=.085$).

Figure 2. Mean percentage of intentional predictions for different agents. The "Overall" bars represent all participants. The "Fast RT" and "Slow RT" bars represent participants who responded relatively quickly and slowly respectively.

The results of this experiment are clearly consistent with the transition model. Participants who delayed their response were less likely to differentiate human and nonhuman behavior. Of course, the primary limit to this experiment is that participants self selected leaving open the possibility that other individual differences aside from choice of reasoning style caused these results. Therefore, in Experiment 2, we employed an experimental design in which half of participants were encouraged to respond based on their first instinct, and half were required to delay their responses and were instructed to consider their answers more deeply.

3. Experiment two

3.1. Method

3.1.1. Participants

Fifty-one subjects completed experiment two (30 female, 21 male, mean age=20, SD=3.8). Subjects were students at Vanderbilt University (n=35) or volunteers from the university's paid research pool (n=16), and received course credit or $5 in exchange for participating. Of these, 25 completed the "open response" condition and 26 completed the "slowed response" condition.

3.1.2. Materials and Procedure

Subjects responded to five scenarios presented on a laptop computer with 12-inch LCD display, or a desktop computer with 15-inch CRT display. Each scenario required subjects to predict the responses of three different agents, including a computer called Yd3, a robot called ASIMO, and a human called John. All scenarios presented each agent in turn, "looking at" and ambiguously "acting upon" a series of objects, as indicated by dynamically appearing and disappearing arrows. Subjects predicted the final action of each agent by choosing from two arrows pointing toward different objects. Three of these scenarios were the same as those in a previous experiment (Levin et al., 2008). The two new scenarios will be discussed below.

Subjects received spoken and written instructions before beginning the experiment. Those in the "open response" condition had no constraints placed on their response time. Accordingly, it was emphasized, "there is no time limit for your responses." Subjects were instructed to use their first "instinct" or "conclusion." Further, subjects were instructed not to hesitate and to "respond as quickly as possible." Those subjects participating in the "slowed response" condition were constrained by the computer program and thereby required to respond only after three seconds had passed. In this condition verbal and written instructions emphasized this time constraint, asked subjects to "think deeply about all relevant factors," and "consider [their] answer carefully."

3.2. Results and discussion

The difference in intentionality between humans and machines (the average of robots and computers) was significantly greater for the free response condition (14.8%) than for the slow-response condition (1.5%; $t(49)=1.728$, one-tailed $p=.045$). The same was true of the differences between the human and computer (fast condition: 16.0%, slow condition: -.07%; $t(49)=1.752$, one-tailed $p=.043$). The difference in human-robot intentionality was nonsignificantly larger in the free response condition (13.6%) than in the slow response condition (3.8%, $t(49)=1.187$, one-tailed $p=.12$).

Within the fast-response condition, the human-machine (e.g. robot+computer) difference ($t(24)=3.201$, $p=.004$), the human-computer difference ($t(24)=2.2667$, $p=.013$), and the human-robot difference were significant ($t(24)=2.527$, $p=.018$; See Figure 3). These differences did not approach significance in the slow response condition ($t's<1$), and in neither condition did the difference between the computer and robot approach

significance (t's<1).

Figure 3. Mean intentionality of predictions for different agents under fast and slow experimental conditions.

4. Discussion

The results of this experiment are consistent with the hypothesis that the basic contrast between the intentionality of human behavior and the nonintentional behavior of computers and robots is the result of first-line reasoning. The specific form of the finding came as a surprise, and, if replicated, could prove quite interesting. If one looks at what changed the most between the fast-response participants and the slow-response participants, it is the intentionality assigned to the human, not the intentionality assigned to the mechanical systems. Therefore, it appears as though participants' deliberations led them away from a simple goal-directed interpretation of human behavior.

There are a number of interesting possible reasons why this occurred. One possibility is that participants began to generate predictions about human behavior that were still goal-directed, but idiosyncratic enough to incorporate unusual location-oriented goals. For example, in the duck-truck scenario, perhaps the slower participants' attention drifted to the salient location-cues in the labeled grid and as a result, they developed the intuition that these would be relevant to the human's goals. On this hypothesis, the primary effect of first-line defaults was to direct attention to typical goals (e.g. object-oriented goals), and that further consideration might lead people to other less salient, but still plausible, goals inherent to the scene. Some indirect support for this possibility comes from previous research exploring people's justifications for their predictions about visual experience. When asked whether they would detect unexpected visual changes in scenes, participants often dramatically overestimate their success, and sometimes justify their predictions by claiming that they would find the specific objects in the scene interesting (Levin et al., 2000; Levin et al., 2002). In a similar vein it is possible to speculate that the deliberative participants in this experiment focused on the identity of the objects and created ad hoc justifications for unusual goal-changes on the part of the human agent. Future versions of these experiments might use response justifications to test these possibilities.

A key result of this study was that first-line reasoning not only distinguished predictions about human and nonhuman entities, but it also was associated with no difference between computers and robots. Thus, the basic default form of reasoning seems to be to discount the anthropomorphism associated with robots, and this does not change as participants think more deeply. This is consistent with our developmental research, which suggests that as children age they become progressively more able to discount anthropomorphic surface features (e.g. appearance and self-initiated movement (Saylor et al., Unpublished results). However, participants who deliberated more extensively also failed to dissociate computers and robots. This finding should be considered in the context of the previous findings reviewed above, in which participants did make more intentional predictions for the robot after they had witnessed the robot

repeatedly looking at objects. Combined, these two findings suggest that the physical anthropomorphism associated with humanoid robots fails to affect participants' first-line predictions, but that it sometimes does, and sometimes does not affect second-line reasoning (if we assume that the previous object-looking experiment result was due to second-line reasoning). One possible difference between the two findings is that the second-line reasoning involved in Experiment 1 was self-generated, whereas the nominally second-line reasoning in the Experiment 2 was driven more by external task demands.

We would like to point out the advantages of relying both on differences in self-selected strategies and experimentally manipulated strategies. Although most experimentalists gravitate to manipulating participants' reasoning, it is important to note that there are significant, and often under-appreciated, advantages to analyzing differences in self-selected strategies. Most important, the fact that participants spontaneously choose to use different strategies makes it likely that the conceptual distinction between the strategies is relevant for real-world behavior. If we had forced participants to decide quickly, their behavior might have been particular to a set of contingencies that are uncommon in real-world decision making. Accordingly, we would advocate that the best approach is to use converging methods employing both the more naturalistic self-selection procedure with the more controlled forced procedures.

5. Conclusion

This experiment successfully differentiated the reasoning of participants who responded relatively quickly from those who responded more slowly. Consistent with the hypothesis that contrasts between human and nonhuman intelligence represent first-line reasoning, participants who responded quickly were more likely to predict that humans would engage in more characteristically intentional behavior than computers. In addition, neither group differentiated robots and computers by predicting more intentional behavior for the robot. These results represent an important step in understanding how people reason about the thinking inherent to different entities, and they suggest that people have available to them a collections of concepts and strategies that must be accounted for if we are to understand how people might respond to interactive and/or anthropomorphic agents in a variety of situations.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0433653 to DTL and MMS.

References

Apperly, I.A., Riggs, K.J., Simpson, A., Chiavarino, C., Samson, D., 2006. Is belief reasoning automatic? Psych. Sci. 17, 841-846.

Baron-Cohen, S., 1995. Mindblindness: An essay on autism and theory of mind. Bradford Books/MIT Press, Cambridge, MA.

Bloom, P., 1997. Intentionality and word learning. Trends in Cognitive Sci. 1, 9-12.

Bruce, A., Nourkabash, I., Simmons, R., 2002. The role of expressiveness and attention in human-robot interaction. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 4138-4122.

Dennett, D., 1991. Consciousness Explained. Little, Brown & Co., Boston, MA.

Dunning, D.A., Perretta, S., 2002. Automaticity and eyewitness accuracy: A 10- to 12-second rule for distinguishing accurate from inaccurate positive identifications. J. of Applied Psych. 87, 951-962.

Gopnik, A., and Wellman, H.M., 1992. Why the child's theory of mind is really a theory. Mind and Lang. 7, 145-171.

Lee, S. L., Kiesler, S., Lau, I. Y., Chiu, C. Y., 2005. Human mental models of humanoid robots. In: Proceedings of the IEEE 2005 International Conference on Robotics and Automation, pp. 2767- 2772.

Leslie, A.M., Friedman, O., German, T.P., 2004. Core mechanisms of theory of mind. Trends in Cognitive Sci. 8, 528-533.

Levin, D.T., In review. Intention and capacity: A dual heuristic framework for visual metaknowledge.

Levin, D.T., Drivdahl, S.B., Momen, N., Beck, M.R., 2002. False predictions about the detectability of unexpected visual changes: The role of beliefs about attention, memory, and the continuity of attended objects in causing change blindness blindness. Consciousness and Cognition. 11, 507-527.

Levin, D.T., Killingsworth, S.S., Saylor, M.M., 2008. Concepts about the capabilities of computers and robots: A test of the scope of adults' theory of mind. In: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, pp. 57-64.

Levin, D.T., Momen, N., Drivdahl, S.B., Simons, D.J., 2000. Change blindness blindness: The metacognitive error of overestimating change-detection ability. Visual Cognition. 7, 397-412.

Levin, D.T., Saylor, M.M., Killingsworth, S.S., Gordon, S., Kawamura, K., In review. Testing the scope of intentional theory of mind in adults: Predictions about the behavior of computers, robots, and people.

Levin, D.T., Saylor, M.M., Varakin, D.M., Gordon, S.M., Kawamura, K., Wilkes, D.M.,

2006. Thinking about thinking in computers, robots, and people. In: Proceedings of the 5th Annual International Conference on Development and Learning, pp. 49-54.

Nass, C., Moon, Y., 2000. Machines and mindlessness: Social responses to computers. J. of Soc. Issues. 56, 81-103.

Saylor, M.M., Somanader, M., Levin, D.T., Kawamura, K., In press. Defying expectations: How do young children deal with hybrids of basic categories? Brit. J. of Developmental Psych.

Searle, J., 1984. Minds, Brains, and Science. Harvard University Press, Cambridge, MA.

Woodward, A.L., 1998. Infants selectively encode the goal object of an actor's reach. Cognition. 69, 1-34.

Vitae

Daniel T. Levin is an associate professor of psychology in the Department of Psychology and Human Development at Vanderbilt University. His research explores the interface between cognition and perception. His work has been supported by grants from the NSF and NIMH. He is the recipient of a Division III APA New Investigator Award (2001), and his research has been frequently cited and described in introductory texts on cognition, perception, and general psychology, and has been covered in the mass media.

Megan Saylor is an assistant professor of psychology at Vanderbilt University. She received her Ph.D. from the University of Oregon in 2001. Her research focuses on the emergence of language and intentional understanding in infants and young children, including young infants' analysis of intentional action, and older children's sophisticated use of cues relevant to referential intentions.

Simon Lynn is a research assistant in the Department of Psychology and Human Development at Vanderbilt University and pursuing a M.A. in sociology from Middle Tennessee State University where he is researching the role of personal beliefs and social norms in extradyadic relationship behavior.

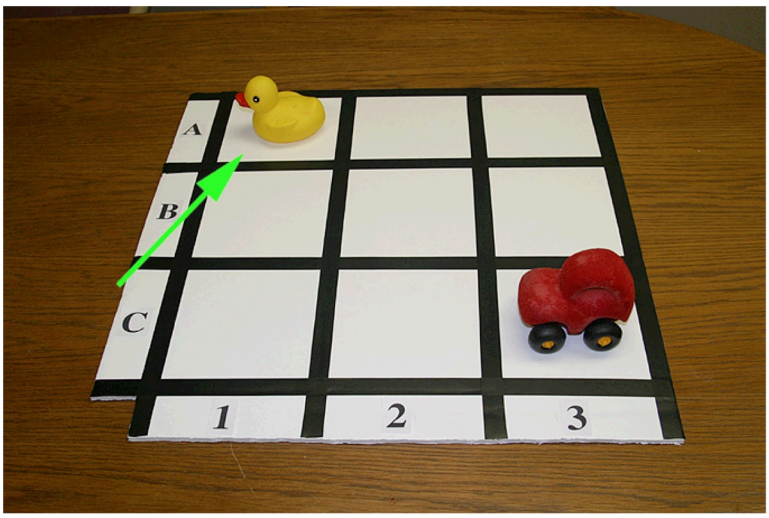

Figure 1

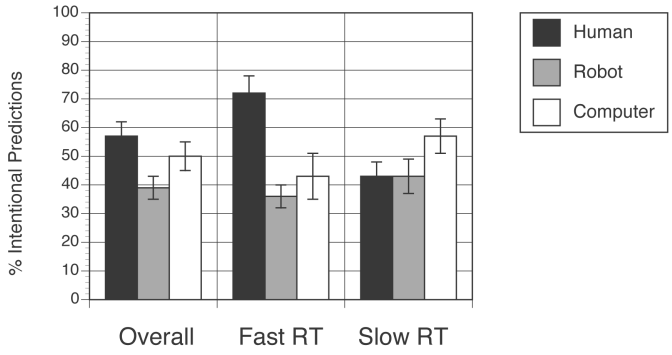

Figure 2

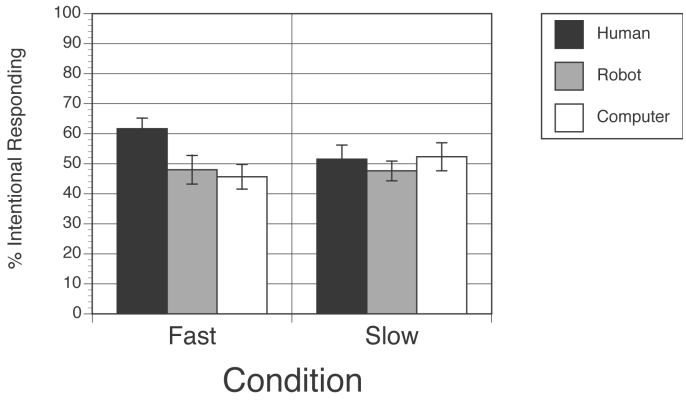

Figure 3