Concepts About the Capabilities of Computers and Robots: A test of the scope of adults' theory of mind

Daniel T. Levin
Vanderbilt University
Dept. Psychology and Human Development
Nashville, TN 37203
daniel.t.levin@vanderbilt.edu

Stephen S. Killingsworth
Vanderbilt University
Dept. Psychology and Human Development
Nashville, TN 37203
stephenkillingsworth@gmail.com

Megan M. Saylor
Vanderbilt University
Dept. Psychology and Human Development
Nashville, TN 37203
m.saylor@vanderbilt.edu

ABSTRACT

We have previously demonstrated that people apply fundamentally different concepts to mechanical agents and human agents, assuming that mechanical agents engage in more location-based, and feature-based behaviors whereas humans engage in more goal-based, and category-based behavior. We also found that attributions about anthropomorphic agents such as robots are very similar to those about computers, unless subjects are asked to attend closely to specific intentional-appearing behaviors. In the present studies, we ask whether subjects initially do not attribute intentionality to robots because they believe that temporary limits in current technology preclude real intelligent behavior. In addition, we ask whether a basic categorization as an artifact affords lessened attributions of intentionality. We find that subjects assume that robots created with future technology may become more intentional, but will not be fully equivalent to humans, and that even a fully human-controlled robot will not be as intentional as a human. These results suggest that subjects strongly distinguish intelligent agents based on intentionality, and that the basic living/mechanical distinction is powerful enough, even in adults, to make it difficult for adults to assent to the possibility that mechanical things can be fully intentional.

Categories and Subject Descriptors: H.1.2 [**Information systems**]: User/machine systems – *human factors.*

General Terms: Experimentation, Human Factors.

Keywords: HRI, Concepts, Theory of Mind.

1. INTRODUCTION

As people interact with an increasing variety of intelligent technologies, they are faced with an array of machines that are, in some ways, very similar to people, and in some ways, very different. This makes it important to understand how people construe the internal processes inherent to artificial minds. Previous research has explored how subjects attribute specific knowledge to intelligent artifacts such as robots [1], and how people are more likely to interact with robots that produce social cues [2]. However, little research has asked whether people believe there are basic, broadly applicable differences between human thought and computerized thought. Moreover, if there are such differences, how will they be applied to humanoid robots that have features of both living things and complex artifacts? In this paper, we describe research that tests the degree to which subjects make broad distinctions among the mental processes inherent to humans, computers, and robots. We then present two experiments testing the degree to which these distinctions are based on beliefs about transitory limits to technology, and ask whether they are caused by a global unwillingness to differentiate among any intelligent artifacts.

One dimension that human and computerized thought is likely to vary along is in the use of intentional representations. Intentional representations are characteristic of human thought, and are special in that they are closely linked to their referents by a network of sensations and knowledge. This means that such representations would be very difficult to "fool" by substituting one referent for another [3]. In contrast, the nonintentional representations characteristic of computers are much less closely linked to their referents, such that one referent could be swapped with another without the system "noticing" anything is wrong. In a sense, one could argue that the nonintentionality of computer representations allows computers to operate on a series of symbols without really knowing what they mean [4].

Closely related to the connectedness of intentional representations is the idea that intentional analysis is reflected in a specific theory about mental functioning that ascribes beliefs, desires, and goals to people, and uses these to explain their behavior [5]. This theory, referred to as an "intentional theory of mind" (TOM), is generally thought to be the mechanism people use to understand other people. Research exploring TOM suggests that over their first 5 years, children move from interpreting others' behavior exclusively in terms of links between external surface behaviors (e.g., eye gaze and objects of focus) to reasoning about internal mental representations (e.g., that looking at an object creates a representation of an object that the person can act on; [6]). The key to this developmental sequence is that children come to explain and predict behavior not with reference to the simple visual cues that precede actions, but rather by hypothesizing that people act based on a series of internal mental representations of beliefs, desires, and goals. This mode of explanation helps children understand how people act based on their *understanding* of the world, not its actual state.

Most research exploring TOM has been done with children, as they begin to understand their own and others' behavior. The few studies that have explored adults' reasoning about others' representation demonstrate that adults continue to struggle in

HRI'08, March 12–15, 2008, Amsterdam, The Netherlands.

understanding others' knowledge and perspective, especially in relatively complex or time-pressured situations [7], [8]. Thus, adults have a more well-established TOM than children, but it is unlikely that they always arrive at clear-cut answers about thinking/representation, or that they always reason effectively about its entailments.

This uncertainty makes it particularly interesting to ask about the range of entities that adults apply TOM to. On one view, adults will apply TOM very broadly, not only to people, but to other living things, and intelligent artifacts as well. Research by Nass and colleagues demonstrates that people treat computers as fully human social actors when interacting with them, for example, feeling obligated to reciprocate, and avoid criticizing them "to their face" [9]. In addition, studies with children and adults show that minimal conceptual and perceptual cues are sometimes sufficient to get subjects to treat actions in graphic displays as intentional and goal-directed [10]. Accordingly, one might predict that adults would presume that any intelligent artifact should be capable of having the same beliefs, desires, and goals as people, or, if they do make a distinction between computers and people, that some simple cues such as the anthropomorphic form of humanoid robots would be sufficient to include them as intentional agents.

On the other hand, another set of findings suggests that both children and adults do strongly distinguish between entities. The most well known developmental findings document this contrast in infants by revealing how they make a fundamental distinction between the goal-directed action of a person, and the location-directed action of a machine [11]. In these experiments, infants are shown a hand, or a mechanical-looking stick repeatedly moving toward one of a pair of adjacent objects, for example a toy duck, and therefore ignoring another object, for example, a toy truck. During the test trials the locations of the objects are switched, and the hand/stick either moves toward the old object in the new location (again reaching for the duck), or reaches toward the new objects in the old location (now reaching for the truck). Results demonstrate that when a hand is doing the reaching, infants look relatively longer when the hand reaches to the new object in the old location, suggesting that they perceive a new goal - the person now wants something new. In contrast, they do not distinguish the machine's movements to the new object or the new location, suggesting that they do not attribute goals to the machine. More generally, authors frequently assume that adults do not believe that computers and people think alike [12], [4], and experimental subjects often report that they would not apply social norms to computers, even when their on-line behavior suggests they do [13]. A similar phenomenon can be observed when studying beliefs about other entities such as God. Many adults explicitly report that they believe that God is not spatiotemporally limited, and does not have an attentional bottleneck, but nonetheless falsely recognize narratives consistent with these assumptions [14].

A subtext in much of this research and theory is that the more implicit, automatic cognitions are a valid measure of subjects "true" beliefs, and that their more explicit beliefs can be discounted as attempts at self-consistent presentation. We would like to argue against this view, and suggest that understanding both kinds of knowledge is important. Implicit beliefs may be important in a wide range of situations, but explicit beliefs have a strong impact on how we solve problems, communicate facts to others, and attempt to deliberate about novel situations. In the present setting, explicit beliefs about human and machine intelligence have been only rarely studied in a systematic way in spite of the fact that they represent a potentially rich set of data about the extension of the adult Theory of Mind. Accordingly, we have asked whether adults make fundamental distinctions among computers, robots, and people by assuming that humans engage in goal-directed actions, while computers, and perhaps robots, do not [15]. We tested this hypothesis by asking adults to make behavioral predictions in a series of scenarios, some of which were modeled on Woodward's infant paradigm. In the scenario most similar to Woodward's, subjects were shown a pair of objects on a labeled grid (see Figure 1), and asked to imagine that each entity had "picked up" the duck at location A1, in each of the first two trials of a three-trial sequence (Figure 1a). Then, subjects were shown a new image with the locations of the duck and truck switched (Figure 1b), and asked what each entity would do on a third trial - would they pick up the duck at the new location, or go back to the old location and pick up the new object. Following Woodward's logic, if subjects are treating the behavior in a goal-directed fashion, they should predict that the entity will retain the same goal on the third trial, and pick up the duck again. On the other hand, if the entity was acting in a more rote (and possibly location-oriented) fashion, subjects should predict that it would reach again to the old location, despite the presence of the new object.

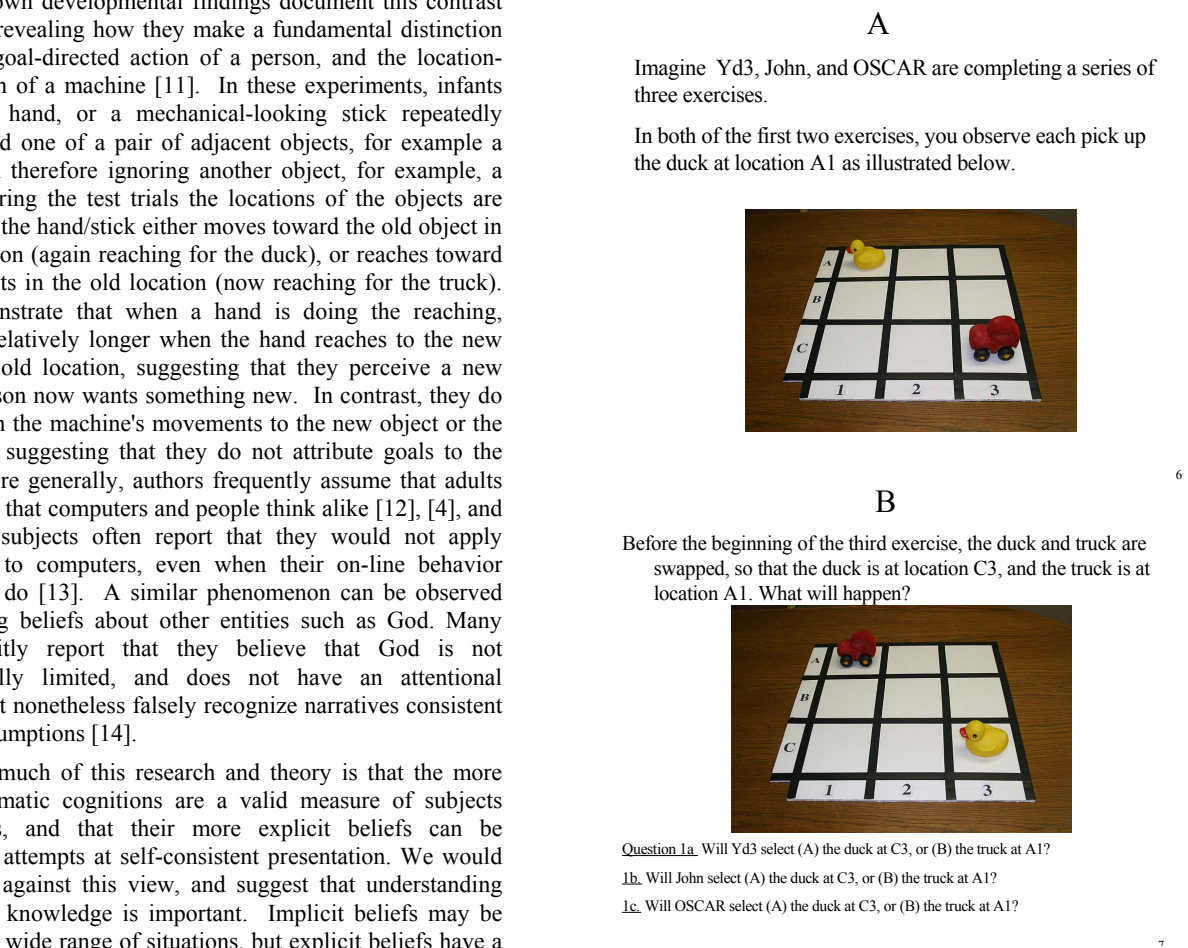

Figure 1. Object vs. Location scenario.

In addition to goal-oriented action, we tested another possible contrast between human and computer thought. That is, the ability to organize the world using taxonomic, or knowledge-based categories, vs. feature-bound/perceptual categories. For example, one might organize the objects depicted in Figure 2 perceptually by placing the large, dark, rectilinear objects, in one category, and the small colorful objects in the other. Alternatively, one could rely on the taxonomic knowledge that some of the objects are food, and some are office supplies, and group accordingly. A wide range of studies in cognition and cognitive development have pitted these two means of categorical organization against each other, and the (often implicit) assumption has been that knowledge-based categories represent a more sophisticated, and deeper grouping than perceptual categorization. Most important for present purposes is the idea that the knowledge underlying many broad artifact categories is inherently intentional in that the commonalities binding these categories are the goals of the humans who created the objects [16]. Accordingly, if subjects believe that an entity is intentional it will choose a taxonomic categorical organization, whereas if it is nonintentional it will organize based on surface features.

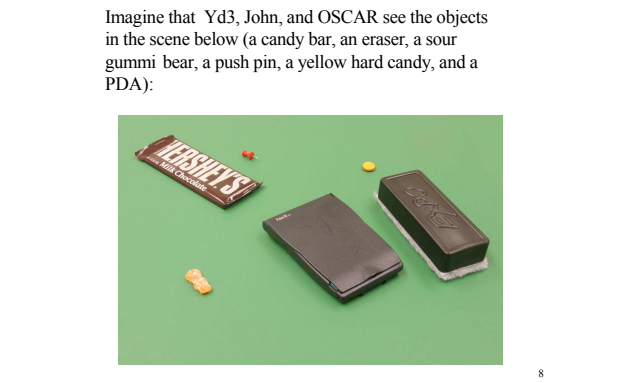

Figure 2. The Feature vs. Category scenario.

Results of our first several experiments demonstrated that adults do, indeed, strongly distinguish computers and people by making far fewer intentional responses (e.g. fewer responses implying goals or taxonomic categorical organization) for computers than for humans [15]. We also tested subjects' predictions about humanoid robots to determine the degree to which their anthropomorphic form would induce more intentional responses (e.g. object/goal-based, and taxonomic) than for computers. In our first experiment we anthropomorphized the robot by simply giving it a human name ("OSCAR"), and describing it as having goals. This description did not lead subjects to differentiate it from computers - they assumed that the robot would produce location-based behaviors and do perceptual classifications just as frequently as the computer. In a second experiment, we added a video showing the robot walking, running, and stopping to let humans pass. This also did nothing to change subjects' attributions. In two additional experiments, however, we asked subjects to track a robot's focus of attention as it was shown pairs of objects. The robot was shown (in a video) looking to one of the objects, and then the other, and subjects were told that they would need to remember which of the two the robot "preferred". For some subjects, this preference was signaled using the duration of the robot's look at each object (under the assumption that the robot would look longer at the preferred object), and for others an "excitement meter" was added (this was a simple graphical display contained in the torso of the robot in which a moving line deflect upwards for preferred objects). After tracking the robot's looking for 10 object pairs, and being tested for memory for the preferences, subjects made their behavioral predictions. In both of two experiments, this manipulation was sufficient to induce more intentional responding for the robot than the computer.

It is important to note that subjects' nonintentioal responses for computers and robots were not simply the result of their belief that current computers are inherently unintelligent. In all of the experiments described above, we asked subjects to rate how intelligent they thought computers were, and to rate how effectively computers can infer human goals based on visual information. In a multiple regression using these ratings, along with subjects' age and sex, to predict the difference in intentionality between humans and computers/robots in each subject's data we found that only ratings of computer goal-understanding, and not general intelligence were predictive. Thus, subjects who believed that computers are good at understanding human goals showed less of a contrast between humans and intelligent machines in behavioral predictions, while there was no such relationship for beliefs about overall intelligence.

This research therefore demonstrates not only a strong contrast in subjects' intuitions about the cognitive processes inherent to different entities, but it also demonstrates that simple anthropomorphism is not sufficient to overcome this difference, at least for the kind of explicit behavioral predictions we tested. Instead, it appears necessary for subjects to carefully track the attentional focus of a robot. It is possible that his tracking is effective because it goes beyond a simple anthropomorphic classification in requiring subjects to practice taking an "intentional stance" in considering the robot's behavior.

Although these previous experiments establish a broadly applicable contrast in people's beliefs about different entities, they leave a number of questions unanswered. First, in light of repeated findings that subjects do not distinguish computers and robots based on simple labels and descriptions, we wanted to verify that subjects would make distinctions in intentional predictions for different intelligent artifacts. It is possible that

subjects simply lump intelligent artifacts into one undifferentiated category, and, until they have some direct experience interacting with them, or tracking their attention, do not consider factors that might allow them to be more intentional. Conversely, we were interested in whether subjects' willingness to differentiate computers/robots from people reflects beliefs in a relatively deep difference in how these systems might think, and not a transitory limit in technology.

2. EXPERIMENT 1

In Experiment 1, we manipulated our description of the robots in two different ways. One group of subjects was asked to give predictions for a robot from 100 years in the future, and subjects the other condition were asked to make predictions about a robot that is remote-controlled by a human. In this experiment, the other entities remained the same. If subjects respond wholly on the basis of their categorization of the robot, independent of temporary technological limits, then we might expect that they will continue to equate the robot and computer, even when comparing a present-day computer to a future robot, and even when comparing a human-controlled robot to a computer. If, on the other hand, the entity category is only a weak reflection of current technological limits, then we would expect that the future robot would not be different at all from the human.

The prediction for the comparison of the human and the human-controlled robot is less clear, but if subjects do not fully equate these, this may be a sign that classification as a robot has "incorrigible" status [10], because an entailment of the classification survives even in the face of logic suggesting that the two should be treated the same.

2.1 Method

2.1.1 Subjects

Fifty-four subjects completed Experiment 1 (26 male, 28 female, mean age=32.6, SD=14.6). Subjects were Vanderbilt Medical Center employees recruited from the hospital cafeteria in exchange for candy (n=27), and students in a general psychology course at Nashville Community College (n=27). Of these, 26 completed the "Future-robot" condition, and 28 completed the "Human-robot" condition.

2.1.2 Materials and Procedure

In this experiment, subjects made behavioral predictions in a series of four scenarios. First, subjects read general directions. The first page of directions informed subjects that we were interested in their "intuitions about three different kinds of things: a person named John, a robot called OSCAR, and a computer system called Yd3." and emphasized that "there are no right or wrong answers - just respond based on your judgment about what each thing will do." This was followed by a description of each entity. In the future-robot condition, the description of OSCAR, the robot asked subjects to "Imagine that OSCAR is a robot built 100 years in the future after many technological advances have been made in robotics". In the human-controlled robot condition, this description read, "OSCAR is a robot that is remote-controlled by a human operator". In addition the means used by the robot and the computer to reach for objects were described: "OSCAR can physically grab objects at different locations using his arm and Yd3 has been loaded into a system that can physically lift objects at different locations using a mechanical vacuum device." We invented the idea of a mechanical vacuum device to avoid the possibility that subjects would anthropomorphize the computer by imagining that it moved things using an arm.

Following the general directions, the three entities were pictured on separate pages. "John" was illustrated with a picture of a White male with a neutral expression, OSCAR was illustrated with a picture of an anthropomorphic robot with arms, a head, and a body, and Yd3 was illustrated with an LCD computer monitor attached to a keyboard and mouse. Underneath the illustration of each entity was an agent-appropriate instruction of the following form: "When making your responses remember that [John, Yd3, OSCAR] is a [human, computer system, human-controlled robot/robot from the future]. Consider what kind of processes characterize a [human, computer system, human-controlled robot/a robot from the future] as opposed to another kind of thing."

The four scenarios were then presented to the subjects. The first scenario (the “Object vs. Location” scenario, see Figure 1) described two trials of a “reaching exercise” in which an entity reached for one of two objects in two locations on a grid. Then, the objects were switched to new locations, and the subject was asked whether the entity would reach to the old location (and therefore the new object) or to the new location (and the old object). The intentional response was to reach for the old object at the new location, demonstrating a belief that the entity was engaged in a goal-directed reach to the object and would reach for it again.

In the “Feature vs. Category” scenario, an array of six objects was pictured first in a disorganized state, then depicted as organized in two different ways. In one organization, the objects were grouped by perceptual similarity (the darker, square objects were grouped), and in another they were organized by semantic category (candy and office supplies). Subjects were asked which of the two organizational schemes the entity would use. The categorical organization would be putatively characteristic of an intentional system.

In the “Position vs. Category” scenario, subjects were told that the entity “reached” for the first, third, and fifth item in a row including writing utensils, and other similarly shaped objects. The three reached-for items were writing utensils, the 6th item was a marker, and the 7th was a screwdriver. The question is whether the system would continue a spatial pattern of reaching and go for the 7th item (the nonintentional response), or continue reaching for writing utensils and reach for the marker at position 6.

In the “Name vs. Location” scenario, subjects were shown a picture of a floppy disk, and a red pen and asked whether it would be better to direct the entity to “lift the red pen” (intentional) or to “lift the object on the left” (mechanical response).

Subjects gave predictions for each entity for each scenario. The order of predictions for the different entities was counterbalanced across subjects. However, the order of the scenarios was the same for all subjects.

After completing the scenarios, subjects completed questionnaires asking them a range of questions regarding their beliefs about, and experience with computers and robots. Two of these questions will be analyzed here: one asking subjects to rate how "intelligent" current computers are, and one asking them to rate the degree to which current computers can "infer the goals of

human action from visual information". An analysis of responses to these two questions by all subjects in Experiments 1 and 2 will be reported in the results section of Experiment 2. In addition, the Nashville Community College subjects were asked a series of open-ended questions regarding their general beliefs about robots and computers. One of these asked whether robots would "think exactly the same way as humans" 100 years in the future. A summary of these responses will be reported in each experiment.

2.2 Results

Although subjects did report that the future robot would produce more intentional behaviors than the current computer (54% vs. 31%; $t(27)=3.855$, $p=.001$), the future robot remained different from the human (54% vs. 79%; $t(27)=4.143$, $p<.001$; See Figure 3). The difference between the computer and the human was also significant, $t(27)=6.681$, $p<.001$. In the Human-controlled robot condition, subjects again predicted more intentional responses for the robot than the computer (61% vs. 35%, $t(25)=4.244$, $p<.001$). In addition, the proportion of intentional responses for the human was higher than that for the human-controlled robot (80% vs. 61%, $t(25)=4.261$, $p<.001$). The difference between the computer and the human was also significant, $t(25)=6.649$, $p<.001$.

Of the 27 subjects asked whether robots would be able to think like humans 100 years in the future, 16 (59%) clearly indicated they would not, 3 indicated they might be able to, 3 indicated they would be able to, and 3 did not respond to the question or gave ambiguous responses. The most frequent justification for a negative response (9 of the 16 negative responses) was that robots would not develop emotions. Four subjects gave some variant of the response that robots are programmed or, conversely, that humans know/learn things on their own. One subject gave each of the following explanations: robots would not be able to "read a person's mind", robots have no soul, robots were not the product of adaptation or evolution, robots have no morality, and robots could produce only "simple motor movements". Note that some subjects gave more than one response to this question.

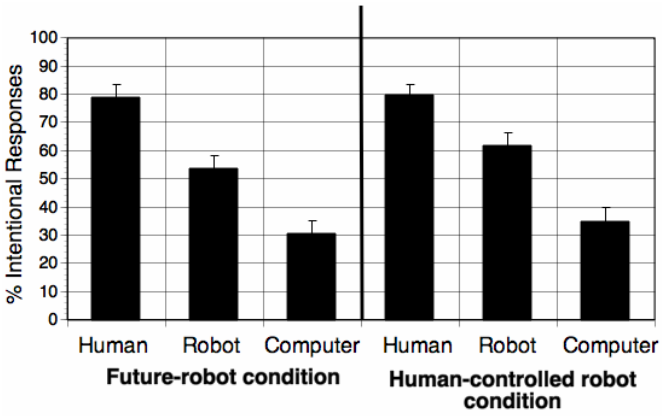

Figure 3. Percentage of intentional responses for Experiment 1. Error bars represent standard errors.

2.3 Discussion

Experiment 1 clearly demonstrates that subjects believe that a future computer will be characterized by some additional intention oriented behaviors, but that time will not entirely erase the difference between these entities. In Experiment 2, we replicated this effect using future versions of all of the entities. Not only should this make a direct comparison between them more valid, but it should also make the concept of future technology more salient. To enhance this salience, we used illustrations of a futuristic computer, and of the robot ASIMO both of which were described as being from 100 years in the future.

3. EXPERIMENT 2

In Experiment 1, we chose to isolate the future robot to allow for a direct comparison between the future robot, and the current technology represented by the current computer. Clearly, this has the disadvantage of forcing a comparison between entities with two differences (e.g. anthropomorphism and future technology), so this issue will be eliminated in Experiment 2, which asks subjects to compare future versions of all of the entities. In particular, we were interested in whether a simple anthropomorphic label effect that had not previously been observed would occur in future versions of a computer and a robot.

3.1 Method

3.1.1 Subjects

Fifteen subjects completed Experiment 4 (6 male, 9 female, mean age=31.6). Subjects were Vanderbilt Medical Center employees recruited from the hospital cafeteria in exchange for candy.

3.1.2 Materials and Procedure

The materials and procedure were very similar to those in Experiment 1 with the following exceptions. First, all three entities were described as being from 100 years in the future on the initial description sheet, and were all described as being from the future on each of sheets depicting the individual entities. The robot and the human were the same as in Experiment 1 (the person in both experiments was pictured in a loose fitting sweatshirt that could plausibly be from almost any time period). The pictured computer was fictional, and had a futuristic appearance, with a very thin looking monitor, and very flat keyboard, and futuristic wireless speakers.

After completing the behavioral predictions, subjects responded to a series of open-ended questions asking them about the capabilities of future robots and computers.

3.2 Results

Subjects gave significantly fewer intentional responses for the future computer than the human (55% vs. 83%, $t(14)=2.915$, $p=.011$), and they gave significantly fewer intentional responses for the robot than for the human (53% vs. 83%, $t(14)=3.055$, $p=.009$). The difference between the computer and the robot was nonsignificant. The future-robot results from this experiment were very similar to those in Experiment 1 - in that case subjects gave 54% intentional responses for the robot, as compared with 53% in the current experiment.

The open-ended questions about the possibility that computers/robots would match human capabilities were consistent with the behavioral predictions in revealing that most subjects believe that computers and robots will not think in the same way as people 100 yeas from now. Ten of 15 (67%) subjects indicated

that a computers or robots (or both) would not think in the same way as humans, three indicated that they might reach the same level as humans, and two indicated that they would. Of the ten subjects who indicated that robots would never be equivalent to humans, four indicated that robots would never have emotions, and one each indicated that humans were too complex to be matched, that robots are programmed while humans are not, that robots could not have empathy, and that robots would have no personality.

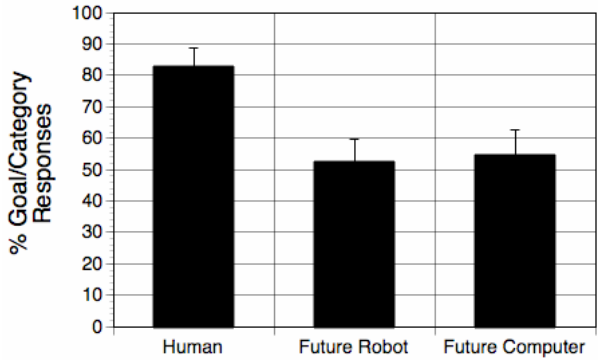

Figure 4. Results of Experiment 2. Error bars are standard errors.

3.3 Discussion

The results of Experiment 2 suggest that subjects predict that the future computer will benefit from advances in technology just as much as the future robot. Therefore, consistent with the two experiments in our previous report [15], we have again observed that simple labels and illustrations do not lead subjects to assume that robots are more intentional than computers. As in Experiment 1, however, subjects do not appear to believe that differences in thinking between people and machines will be erased, at least in the foreseeable future. This intuition is reflected in responses on the post-experiment questionnaire. When asked whether computers or robots will think in the same way 100 years in the future most subjects indicated it would not.

4. GENERAL DISCUSSION

The data reported here, combined with that reported in our pervious experiments demonstrate that people make robust distinctions among different intelligent entities by assuming that human actions are founded upon goals, and knowledge organized taxonomic categories, while computers are more location oriented, and focus more on perceptual features as a means of organizing object categories. In this section we discuss some implications of these findings for understanding people's concepts about the cognition inherent to different entities. First, however, it is important to consider one potential interpretive issue with both of these experiments.

Although subjects clearly differentiated the future robot, and the human-controlled robots from the current computer they did not go so far as to equate either to the human. This is particularly striking in the case of the human-controlled robot, which might be expected to make precisely the same decisions as a person. This result suggests that subjects presume a deep difference between human intelligence and machine intelligence, but the possibility that this difference occurred because some subjects did not consistently apply the experimental instructions must also be considered. If a subset of subjects did not remember that they were supposed to make predictions for a future robot, or a human-controlled robot, then the number of intentional judgments might be artificially low. Although this possibility cannot be completely ruled out, several findings mitigate against it. First, the salience of the future-manipulation in Experiment 2 was much greater than that in Experiment 1: because all three entities were from the future in Experiment 2 (as opposed to just the future-robot in Experiment 1), and because a future-computer illustration was added, the experimental context provided much stronger set of reminders that subjects were rating future entities. In spite of this difference, the proportion of intentional ratings for the future robot in both experiments was almost exactly the same (54% in Experiment 1, and 53% in Experiment 2). If subjects were responding simply based on the availability of the manipulation, one would expect that subjects would attribute more intentional responding to the future-robot in Experiment 2. In addition, the open-ended responses in Experiment 2 confirmed that many subjects were willing to explicitly state that computers and robots would not think the same as people in the foreseeable future.

Combined, the behavioral predictions and the open-ended responses converge to demonstrate that people believe that there are deep differences in the basic kind of thinking done by computers, robots, and humans that are not simply consequences of transitory limits in current technology. The depth of this difference, and the possibility that it is closely tied with an ingrained living-nonliving distinction, are reinforced by the finding that human-controlled robots are not seen as fully intentional. Neither, however, are subjects entirely inflexible in considering the possibility that the difference between human and machine intelligence will lessen over time.

Given the strong contrast in predicted performance between artificial entities and humans, it is important to consider what differentiates these findings from others that find more equivalence. One possibility is that equivalence is observed when subjects do not explicitly consider how computers and robot think, and instead rely on more heuristic/associative reasoning [17]. For example, the well-known findings bv Nass and colleagues probably reflect the activation of relatively automatic social schemas and scripts whereas our scenarios ask subjects to directly consider the system's capabilities. Although this contrast can help understand some results, it needs to be more refined to fully capture the pattern of results we have observed here and elsewhere. In several different experiments we have observed that anthropomorphism can sometimes induce intentional predictions. As mentioned in the introduction, these manipulations seem to require more than simple labeling – in one set of previous studies they required subjects to attend to a robot focusing attention on a series of objects [14], and in another previous study it required subjects to learn the details of a machine vision system in an anthropomorphic context [18]. Both of these situations require that subjects interpret visual behavior, or visual processing using an intentional context to support explicit reasoning.

This pattern of results could mean that subjects have an explicit set of concepts that strongly and categorically delineates nonintenional machine intelligence from intentional human intelligence, and another implicit set of concepts that attributes intentionality to any of a wide range of systems so long as they

have some minimal, characteristic signs of intentionality. This more implicit system might reflect the early-emerging distinctions observed by developmental psychologists, whereas more explicit concepts that equate all machine intelligence may emerge later. One potentially problematic aspect of this distinction is the hypothesis that the more expansive intentionality is necessarily implicit. People are probably aware of their own anthropomorphization of many machines, so it might be better to consider the possibility that people can select among several models to use in helping them understand machine behavior. In addition to applying TOM, or a purely mechanical approach subjects might, for example, also apply a much more limited similarity-based theory about intelligent artifacts, making inferences about their behavior based solely on previous experience with very similar actions. This kind of reasoning probably most prominent in familiar settings where individual use specific commands to achieve specific tasks. The more broad theory-based approaches might be more relevant in novel settings where subjects must make sense of a new technology, or must explain unexpected behavior in a more familiar setting.

These findings demonstrate that subjects have clear beliefs about how mechanical agents think, and that these beliefs are responsive to some, but not all, anthropomorphic contexts. It is important to note that these beliefs are apparently not isolated. In other research, we have found that subjects' attributions about action perception also differentiate computers and humans. In particular, subjects assume that computers process action in smaller "chunks" than people, and reveal this assumption by segmenting continuous action streams into smaller units for a computer audience [19]. As in the current experiments, this contrast is predicted by ratings of computers' ability to understand human goals. In addition, we have found that people produce actions differently for computer and human audiences, emphasizing small units with more clear movements for computers, and producing more communicative gestures for humans [20]. Combined, these results help delineate people's knowledge about the behavior and internal processes inherent to different intelligent entities. If we can understand these, it may be possible not only to improve people's interactions with intelligent machines, but also guide machine behavior to be consistent with people's assumptions about different kinds of intelligence.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] 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 (Barcelona, Spain, April, 2003). 2767- 2772

[2] Bruce, A., Nourbakhsh, I. and Simmons, R. 2002. The role of expressiveness and attention in human-robot interaction. In Proceedings of the IEEE International Conference on Robotics and Automation (May, 2002), 4138--4142, May 2002.

[3] Dennett, D. 1991. Consciousness Explained, Boston, MA: Little, Brown & Co.

[4] Searle, J. 1984. Minds, Brains, and Science. Cambridge, MA: Harvard University Press.

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

[6] Gopnik, A., Slaughter, V., and Meltzoff, A. 1994. Changing your views: How understanding visual perception can lead to a new theory of the mind. In Origins of an understanding of mind, C. Lewis & P. Mitchell Eds. Hillsdale, N.J.: Erlbaum, 157-181.

[7] Barr, D. and Keysar, B. 2005. Mindreading in an Exotic Case: The Normal Adult Human. In Other Minds: How Humans Bridge the Divide between Self and Others, B. F. Malle and S. D. Hodges Eds. New York, Guilford Press, 271-283.

[8] Levin, D.T., and Beck, M.R. 2004. Thinking about seeing: Spanning the difference between metacognitive failure and success. In Thinking and Seeing: Visual Metacognition in Adults and Children, D.T. Levin Ed. Cambridge MA: MIT Press.

[9] Nass, C., & Moon, Y. 2000. Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56, 81-103.

[10] Johnson, S.C. 2003. Detecting agents. Phil. Trans. R. Soc. Lond. B., 358, 549-559.

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

[12] Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., and Nass, C.I. 2003. Syntactic alignment between computers and people: The role of belief about mental states. Proceedings of the 25th Annual Conference of the Cognitive Science Society (Boston, MA, July 2003).

[13] Nass, C., Moon, Y., & Carney, P. 1993. Are respondents polite to computers? Social desirability and direct responses to computers. Journal of Applied Social Psychology, 29, 1093-1110.

[14] Barrett, J. L. and Keil, F.C. 1996. Conceptualizing a non-natural entity: Anthropomorphism in God Concepts. Cognitive Psychology, 31, 219-247.

[15] Levin, D.T., Saylor, M.M., Killingsworth, S.K., Gordon, S., & Kawamura, K. in prep. Predictions about the behavior of computers, robots, and people: How does intentionality affect what people think something will do?

[16] Bloom, P. 1997. Intentionality and word learning. Trends in Cognitive Sciences, 1, 9-12.

[17] Sloman, S.A. 1996. The empirical case for two systems of reasoning. Psychological Bulletin, 119, 3-22.

[18] Levin, D.T. in prep. Intention and Capacity: A dual heuristic framework for visual metaknowledge.

[19] Killingsworth, S.S., Saylor, M.M., & Levin, D.T. in review. Intentional understanding through a machine's eyes.

[20] Herberg, J.S., Saylor, M.M., Ratanaswasd, P, Levin, D.T., & Wilkes, D.M. in review. Audience-contingent variation in action demonstrations for humans and computers.