CHAPTER 3

You, Robot

BRIAN FIALA, ADAM ARICO, AND SHAUN NICHOLS

Abstract

How do people think about the mental states of robots? Experimental philosophers have developed various models aiming to specify the factors that drive people's attributions of mental states to robots. Here we report on a new experiment involving robots, the results of which tell against competing models. We advocate a view on which attributions of mental states to robots are driven by the same dual-process architecture that subserves attributions of mental states more generally. In support of this view, we leverage recent psychological research on human-robot-interaction that involves ecologically valid stimuli such as Roombas and humanoid robots.

1. Introduction

The study of folk psychology has recently been reinvigorated by a series of empirical studies investigating the attribution of experiential states to others. These studies have examined not only how people attribute experiential states to other humans, but also to nonhuman entities such as robots.

In one such study, Heather Gray, Kurt Gray, and Dan Wegner (2007) investigated whether everyday mind perception occurs along a single dimension or along multiple dimensions. Do people attribute certain mental states (such as beliefs or intentions) but not others (e.g., experiences of pleasure or hunger) to different types of entities? Gray and colleagues instructed participants to complete a series of pairwise comparisons between various characters, including humans, animals, and a robot. Subjects considered a series of mental capacities (e.g., capacity to feel pain; capacity for planning) and judged which character in each pair best exemplified the mental capacity. According to Gray and colleagues, people attributed mental states along two dissociable dimensions: *agency* and *experience*. One piece of evidence for this conclusion was that people attributed mental capacities to robots along the agency dimension (e.g., memory, planning, and thought), but not the experience dimension (e.g., fear, pain, pleasure).

In their article "Two Conceptions of Subjective Experience," Justin Sytsma and Edouard Machery (2010) similarly defend a view in which folk mind attributions can be carved into two dimensions. However, they propose an alternative way of distinguishing the dimensions, arguing that we should reject the idea that folk psychology represents the philosopher's distinction between *intentional* and *phenomenal* states. Instead, they hold that we can best understand folk psychology as distinguishing between mental states based on whether they possess a "hedonic value," or valence. Some mental states are pleasant or unpleasant (i.e., positively or negatively valenced), whereas other mental states have no such pleasant/unpleasant feature (i.e., unvalenced).[1] According to Sytsma and Machery, the mental states people resist attributing to robots are not phenomenal states generally, but rather states with *hedonic value* specifically. These conclusions are derived from data from a series of empirical studies involving Jimmy the robot:

> Jimmy . . . is a relatively simple robot built at a state university. He has a video camera for eyes, wheels for moving about, and two grasping arms with touch sensors that he can move objects with. As part of a psychological experiment, he was put in a room that was empty except for one blue box, one red box, and one green box (the boxes were identical in all respects except color). An instruction was then transmitted to Jimmy. It read: "Put the red box in front of the door." (Sytsma & Machery, 2010, p. 306)

In one version of the story, Jimmy moves the box in front of the door with no noticeable difficulty. In another version, Jimmy is given an electrical shock when he picks up the box, at which point he drops the box, moves away, and fails to follow subsequent instructions to move the box. Sytsma and Machery surveyed both philosophers and nonphilosophers as to whether Jimmy (in the first story) saw red and (in the second story) felt pain. Whereas philosophers refrained from ascribing either phenomenal state to Jimmy, nonphilosopher participants were happy to say that Jimmy saw red, but abstained from saying that Jimmy felt pain. The fact that philosophers and nonphilosophers presented different response patterns, according to Sytsma and Machery, reveals that folk psychology does not include the same distinction between intentional and phenomenal states that philosophers of the mind standardly employ.

In order to determine more precisely how nonphilosophers come to attribute various states of subjective experience, Sytsma and Machery (2010) then presented subjects with a series of additional vignettes about Jimmy. In one, Jimmy is described as also possessing a scent detector, and the three color boxes from the original story are replaced by three (closed) boxes containing bananas, chocolate, or peeled oranges; mirroring the original story, Jimmy is instructed to place the box containing bananas in front of the door, which he does without difficulty. Participants were then asked, "Did Jimmy smell banana?" In another version, Jimmy is placed in the original room with three colored boxes and is instructed to place the red box in front of the door; however, there is another robot in the room that runs into Jimmy and prevents Jimmy from reaching the red box. Jimmy eventually rams the other robot. When the other robot moves away, Jimmy chases it around the room. Participants were then asked, "Did Jimmy feel anger?" Finally, a scenario is described in which three chemical compounds are placed, one-by-one, under Jimmy's scent detector; Jimmy is then placed in a room containing three boxes, each with one of the chemical compounds, and is instructed to place the box containing isoamyl acetate in front of the door, which he does without noticeable difficulty. Participants were then asked, "Did Jimmy smell Isoamyl Acetate?"[2]

Sytsma and Machery (2010) found that participants were ambivalent about both Jimmy smelling banana and Jimmy feeling anger but were more than willing to say that Jimmy smelled isoamyl acetate. Their explanation is that people associate smelling banana with a positive valence and associate feeling anger with a negative valence, but people have no such expectations for the unfamiliar compound isoamyl acetate. Because smelling banana and feeling anger involve hedonic value, Sytsma and Machery's account predicts that people will not be inclined to attribute these states to a simple robot. Because people do not associate any hedonic property with isoamyl acetate, they have no problem saying that Jimmy can smell it.

Wesley Buckwalter and Mark Phelan (2013) have challenged Sytsma and Machery's (2010) positive thesis that folk attributions of subjective experience are based on considerations of hedonic value. Instead, they argue that what people are really responding to in the vignettes is Jimmy's function, in some broadly teleological sense. According to Buckwalter and Phelan, people's mental state attributions are driven by "tacit assumptions on the part of experimental participants about the function for which Jimmy was created."[3] The Sytsma and Machery results, then, would not seem to reveal a hedonic-based categorization of subjective experience in folk psychology but rather reflect the differences in Jimmy's intended function across the various vignettes.

To test this hypothesis, Buckwalter and Phelan (2013) devised Jimmy vignettes of their own. In one version, Jimmy—outfitted as in the original vignettes with video camera eyes, scent detector, wheels, and grasping arms—was created for the purpose of cleaning biomedical waste; in another, Jimmy—outfitted identically—was designed for the purpose of making fruit smoothies. In both conditions, participants were then asked, "Did Jimmy smell vomit?" and "Did Jimmy smell banana?" Buckwalter and Phelan report that participants in the biomedical waste condition were significantly more likely to say that Jimmy smelled vomit than were participants in the smoothie condition, and participants in the smoothie condition were more likely to say that Jimmy smelled bananas than were participants in the biomedical waste condition.

Buckwalter and Phelan take these data to confirm their thesis that people attribute mental states based on the functional roles that the system is designed to realize. On this view, we can understand the original Sytsma and Machery (2010) results might be due to the fact that the original vignettes suggest that Jimmy was designed for the purpose of visually identifying and moving particularly colored boxes and not for pain-related tasks.

Bryce Huebner (2010) has offered an account that is similar to Buckwalter and Phelan's (2013) view, insofar as both appeal to a notion of function in explaining people's mental state attributions. Whereas Buckwalter and Phelan emphasize "function" in the sense of design purpose, Huebner focuses on "function" qua causal role. He set out to answer the question, "Do people rely on the structural [i.e., implementational] properties of an entity in making judgments about the acceptability of a mental state ascription, or are they more concerned with that entity's functional organization?" To this end, Huebner probed subjects to find out whether they would endorse various mental state attributions to four different entities: a normal human, a human with a computer central processing unit (CPU) in place of a brain, a robot with a brain in place of a CPU, and a robot with a CPU. He found that while participants voiced agreement with attributions of non-phenomenal states (such as beliefs) to all four entities, participants were far less generous with phenomenal states (such as feeling pain or feeling happy). The majority of participants said that the ordinary human could feel pain (82%) and feel happy (76%), and the majority said the ordinary robot neither felt pain (56%) nor felt happy (% not reported), but responses were "essentially at chance" for both of the "cyborg" cases (human with computer CPU and robot with human brain). Huebner concludes that folk psychology seems to include something like the philosophical concept of non-phenomenal mental states and that attributions of these states seem to rely primarily on functional considerations. However, he also concludes that folk psychology *does not* employ a concept that resembles the philosophical concept of phenomenally consciousness mental states. Folk attributions of such experiential states, according to Huebner, are not treated uniformly, nor are they based on either physiological or functional considerations, but instead rely on considerations of agency and personhood.[4]

All of these experimental results seem to present a problem for the Agency Model of mental state attribution, which we have defended in previous papers (Arico, Fiala, Goldberg, & Nichols, 2011; Fiala, Arico, & Nichols, 2012). In particular, it seems problematic that participants in these studies attributed intentional states to robots but withheld some basic phenomenal state attributions. According to the Agency Model, mental state attributions are governed by a dual-process cognitive system.[5] Although the high-road process operates via slow, conscious, domain-general deliberation, the low-road process operates in a quick, automatic, domain-specific way. The low-road disposition to attribute mental states is the result of categorizing an entity as an AGENT, which is itself a consequence of representing that entity as possessing particular properties: facial features, interactive behavior, or moving in a distinctive trajectory.[6] The Agency Model claims that categorizing an entity as an AGENT is sufficient to generate a disposition to attribute a *wide range* of mental states to the target, including phenomenal states like pain. Because robots such as Jimmy have some of the cue properties, the Agency Model predicts that people should be disposed to attribute phenomenal states to robots.

Often, displaying just one of the triggering cue-properties is sufficient to dispose subjects to attribute mental states. For example, subjects in Heider and Simmel's (1944) experiment attributed mental states such as "wanting," "chasing," and "helping" to the moving geometric figures in a short animation. Because the figures were simple (e.g., triangle, square) and lacked other distinguishing features, it must have been the shapes' distinctive motion trajectories that elicited the wide range of mental state attributions from the subjects. Regarding contingently interactive behavior, Johnson (2003) manipulated interactivity by showing infants a football-shaped, beeping "blob" in one of two conditions: In the first condition, the blob beeped at a confederate at random intervals, emitting a predetermined amount of overall beeping; in the second condition, the blob emitted the same overall amount of beeping but "waited its turn" so that it appeared to have a "conversation" with the confederate. Johnson found that when the blob reoriented its direction, infants were more likely to follow its "gaze" in the interactive behavior condition. Adding facial features to an object also triggers mentalistic attributions. Using a similar gaze-following paradigm, Johnson, Slaughter, and Carey (1998) found that infants were more likely to follow the "gaze" of a fuzzy brown object when eyes were placed on the object. Turning to the range of mental states elicited by these cue features, reaction-time results from Arico et al. (2011) suggest that simple triggers suffice to dispose subjects to attribute pain, a state that is both phenomenal (i.e., there's like something to undergo it) and valenced (i.e., it has a negative hedonic value). Subjects read pairs of words describing a type of object and a type of property, and were asked to judge as quickly as possible whether that kind of object could have that type of property. Subjects showed significantly slower reaction times when judging that objects possessing the cue features (such as ants) could *not* feel pain and relatively quick reaction times when denying that objects such as trucks, clouds, and rivers could feel pain. This suggests that it takes subjects some time to override the fast, automatic inclination to make pain-attributions resulting from representing the target as an AGENT.[7]

The Agency Model claims that triggering the low-road process and categorizing an entity as an AGENT disposes us to attribute both intentional and phenomenal mental states, as well as both valenced and unvalenced mental states. That is to say, low-road processing automatically facilitates a wide range of mental state attributions. The Agency Model does not claim that we would be disposed to attribute *every* kind of mental state, however. Some states (e.g., Schadenfreude or agoraphobia) might be too complex or unusual to be poised for attribution. However, for extremely basic and common mental states such as pain, anger, and sadness, the Agency Model certainly suggests that we should expect people to be disposed to attribute those states to things that have been categorized as AGENTs. As a result, the data discussed earlier seem to present a problem for the Agency Model, because subjects seem willing to attribute to robots only a circumscribed range of mental states. Robots often possess cues such as facial features, distinctive motion trajectories, and contingently interactive behavior. Jimmy possesses at least two of these features, so the Agency Model predicts that subjects should be inclined to attribute all sorts of basic mental states to robots like Jimmy. You might think that champions of the Agency Model should be embarrassed by this state of affairs. Instead, we are eager to demonstrate that the apparent evidence against the Agency Model is merely apparent. The data on robots present an opportunity to elaborate further details of the Agency Model, and we think explaining these data will ultimately make our case for the Agency account stronger rather than undermine it.

One line of response begins with the observation that in the experiments just canvassed, the stimuli were vignettes conveyed primarily by linguistic representations. It is natural to infer from this that no cue features were present in the stimuli and, thus, that the Agency Model should not predict any automatic inclination to attribute mental states. Although this move is tempting, it does not yield a workable defense of the Agency Model. One problem is that although cue features are not strictly present in text-based stimuli, representations of the cues may nonetheless be involved in processing the text (e.g., via semantic associations). Indeed, Arico et al. (2011) presuppose something along these lines, because they obtained an effect in their reaction-time study by presenting subjects with linguistic representations of the object and property categories. The cue features may also be represented in mental imagery. Wheatley, Milleville, and Martin (2007) found that many of the same neural systems are implicated in both perceiving the agent-like motion of simple objects and imagining the same kinds of motion trajectories. Similarly, the task of imagining a face preferentially activates many of the brain areas known to be important for the perception of faces (O'Craven & Kanwisher, 2000). In all of these cases, it is plausible that cue features such as faces or distinctive motion trajectories are represented in subjects, although the cues are not overtly present in the stimuli. Thus, we cannot rest a defense of the Agency Model on the claim that the apparent counterevidence is the product of text-based vignette studies.

A more important consideration is that subjects in vignette studies have an opportunity to spend some time engaging in conscious, high-road reflection before making their judgments about robots. Consequently, we should expect that subjects bring some of their background beliefs to bear and that their judgments are not wholly the product of low-road processing. The Agency Model predicts that when subjects read vignettes about robots, they will typically represent some cue features and thus undergo some inclination to judge that the robot has a wide range of mental states. But the Agency Model does not predict that subjects will overtly judge that robots are conscious, because high-road reflection may cause subjects to override their intuitive inclinations. It is effectively a platitude in our culture that robots are incapable of pain or emotion. Given the cultural prevalence of that attitude, it is reasonable to hypothesize that this belief will figure in high-road reasoning about robots. If so, then subjects will show significant resistance to attributions of mental states to robots generally.

To evaluate attitudes about robot mentality, it's important to distinguish deliberative, high-road responses from automatic low-road responses. We first explore the high road further, building on the vignette studies of previous work. We then turn to the low road, reviewing a diverse body of work exploring how people react to actual computers and robots (as opposed to vignettes about robots).

2. Robots on the High Road

We suspect the design of existing vignette studies on robots puts undue pressure on subjects to attribute certain states to robots. Subjects in these studies face a forced choice and have no way of describing Jimmy's information-processing behavior besides adverting to mental states. We predict that subjects will tend to deny that robots can have a wide range of mental states, if given the opportunity to otherwise communicate this information. To test this prediction, we designed a vignette study that allowed subjects the option of denying that the robot has an unvalenced mental state (such as "seeing") while also acknowledging that the robot is carrying out some relevant function for which it was designed (such as "detecting" or "identifying").

We presented participants with the classic Sytsma and Machery (2010) vignette about Jimmy the robot. But rather than giving them a forced choice, we allowed them to select any descriptions of the robot that seemed right from a set of candidate descriptions (á la Guglielmo & Malle, 2010). The vignette runs as follows:

> Jimmy . . . is a relatively simple robot built at a state university. He has a video camera for eyes, wheels for moving about, and two grasping arms with touch sensors that he can move objects with. As part of a psychological experiment, he was put in a room that was empty except for one blue box, one red box, and one green box (the boxes were identical in all respects except color). An instruction was then transmitted to Jimmy. It read: "Put the green box in front of the door." Jimmy did this with no noticeable difficulty.[8]

After being presented with the vignette and the picture of the robot, participants were asked, "Which of the following descriptions of Jimmy are correct? Check any description that seems right to you." There were five descriptions in fixed order: "Jimmy detected green," "Jimmy saw green," "Jimmy located the green box," "Jimmy identified the green box," and "Jimmy moved the red box." The last item was used as a manipulation check—participants who indicated that this description was correct were excluded.

After excluding those who failed the materials check, we found that only 7 of 25 participants indicated that "Jimmy saw green" was an appropriate description. This is significantly lower than what would be expected by chance alone (χ^2 goodness of fit $= 4.840, p = 0.0278$). Again, following Sytsma and Machery (2010), we also presented participants with a vignette involving a human (their "Timmy" vignette). Participants were presented with the same set of options. In this case, they were more likely to select "Timmy saw green" as a correct description than they were in the robot case ($\chi^2 = 4.567, p = 0.0326$; see Table 1 for responses).

Table 1. Attributions of "saw green"

	Selected "saw"	*Didn't select "saw"*	*Total*
Robot	7	18	25
Human	16	12	28
Total	23	30	53

These results suggest that the overly narrow choice format might artificially inflate the attributions of perceptual experience to robots in Sytsma and Machery's (2010) studies. Participants want to be able to communicate that the robot is processing some information about the environment, mediated by a camera. And because the only question is, "Did Jimmy see green?" the only way to communicate this is by saying yes. When given a chance to communicate more precisely what they think is going on with Jimmy, participants prefer expressions such as "Jimmy detected green" or "Jimmy located the green box."[9]

This result nonetheless leaves open the possibility that subjects are resistant to attributing specifically phenomenal states to robots; it may be that the low ratings on "saw" derive from the fact that "saw" is (arguably) a phenomenal attribution. So we tried another case using a canonically intentional predicate—*knows*. We used the same format as before, presenting participants with the Sytsma and Machery (2010) vignette and then asking them to indicate which descriptions are correct. In this case, the list of descriptions included "Jimmy processes the location of the green box," "Jimmy knows the location of the green box," "Jimmy can detect the location of the green box." "Jimmy identifies the green box," and "Jimmy moved the red box." Once again, we used the last item as a manipulation check. After excluding participants who failed the check, we found that fewer than half of the participants (13 of 27) selected "Jimmy knows . . ." as a correct description. This did not differ from the responses on the "saw" case ($\chi^2 = 2.226$, $p = 0.1357$). We also compared responses on the robot case with a parallel human version of the vignette, and we found that people were (marginally) more likely to select the description "Timmy knows . . ." (21 of 29) in the human case than they were in the robot case ($\chi^2 = 3.452$, $p = 0.0632$; see Table 2 for all responses).

As before, when given a chance to communicate more precisely what they think is going on with Jimmy, participants prefer expressions such as "Jimmy processes the location of the green box" and "Jimmy can detect the location of the green box."[10]

Thus, in these experiments that offer more choices for describing Jimmy's behavior, we do not find that participants are prone to attribute either phenomenal or non-phenomenal mental states to robots. We think there is a straightforward explanation for this—there is a platitude (at least in our

Table 2. Attributions of "knows the location of the green box"

	Selected "knows"	*Didn't select "knows"*	*Total*
Robot	13	14	27
Human	21	8	29
Total	34	22	56

culture) that robots do not have minds. This platitude guides attributions on the high road, leading to attenuated attributions of mental states. The idea that performance on these tasks depends on high-road responses also helps explain why there is so much variance on the tasks. Indeed, in our own experiments, we find that a sizable minority of people are reluctant to attribute "saw" and "knows" even in the human case. This indicates that there is a fair amount of metacognition going on in these cases.

3. Robots on the Low Road

In the previous section, we found that when given a range of options, people tend to prefer nonmental state attributions to robots over mental state attributions. This seems to be an embarrassment for our view in that there is a low-level inclination to attribute a wide range of mental states to objects that are categorized as AGENTs. We have suggested that the responses we see in the vignette studies are typically the product of high-road processing, and this provides a way of rendering them consistent with the Agency Model. After all, the model predicts the low-level inclinations to attribute mental states but allows that those inclinations can potentially be moderated by competing conceptual associations or overridden by higher level cognition. However, for this line of defense to hold, we also need to show that there really is a low-road tendency to attribute a wide range of mental states to robots. That is what we aim to do in the present section. Fortunately, there is a wealth of work inspired by interest in human-computer interaction.

In presenting the Agency Model, we noted that attributions of AGENCY are triggered by distinctive motion trajectories (e.g., Heider & Simmel, 1944), the presence of a face (Johnson et al., 1998), and contingent interaction (Johnson, 2003; Johnson et al., 1998). Of course, we do not think that these cues are the *only* ways that people come to attribute AGENCY. But we organize our coverage around these three cues.

3.1 Motion Trajectories

In a classic experiment on attributions of animacy, Tremoulet and Feldman (2000) had participants observe a small geometric object move on an otherwise blank screen. Participants were told to indicate the degree to which the object is "alive" under a wide variety of motion trajectories. Tremoulet and Feldman found that the convergence of two key cues greatly augmented attribution of animacy: change in speed and change in direction. For instance, if the object turns and speeds up, this leads to high attributions of animacy.[11]

Saerbeck and Bartneck (2010) drew on this work and programmed two robots, a Roomba and an iCat, to exhibit a range of motion trajectories. The Roomba is a commercial robotic vacuum cleaner that looks like a fat Frisbee.

The iCat is a research robot designed for interaction with humans; it has a mechanical face and is shaped like a cat. Participants observed each robot as it exhibited a range of motion patterns. For each motion sequence, participants indicated on a standard pictorial measure of emotions (the SAM scale) which picture best described the behavior of the robot (Bradley & Lang, 1994).[12]

The research showed no difference in attributions for the iCat robot and the Roomba. What did matter, however, were motion trajectories. Saerbeck and Bartneck (2010) found that varying the acceleration and direction affected the attribution of emotion. Changing the acceleration significantly affected attributions of degree of arousal, and changing the direction significantly affected attributions of valence (Saerbeck & Bartneck, 2010, pp. 58–59). There was also a significant interaction between acceleration and direction change on attributions of valence (Saerbeck & Bartneck, 2010, p. 59). In addition, after the main part of the task, when participants were ask to describe the behaviors, "all participants used emotional adjectives to describe the robots' behavior" (Saerbeck & Bartneck, 2010, p. 58). Thus, this provides some reason to think that distinctive motion cues associated with AGENT categorization also facilitate attribution of emotions to robots. It is natural to interpret these emotion attributions as valenced, and they have no obvious connection to the function for which the robots were designed.

3.2 Faces

Are robot "faces" sufficiently human-like to trigger AGENT categorization? Our face-detection system seems to have a hair trigger. The mere presence of googly eyes, as in Johnson et al. (1998), is sufficient to result in categorization as a face, and consequently categorization as an AGENT. Similarly, when adult subjects viewed "schematic faces" composed of simple elements corresponding to eyes, mouth, and nose (e.g., a grounded electrical outlet), those subjects overwhelmingly judged that the object "looks like a face," and displayed increased activation of the fusiform face area (Hadjikhani, Kveraga, Paulami, & Ahlfors, 2009). Tong, Nakayama, Moscovitch, Weinrib, and Kanwisher (2000) also found that the fusiform face area showed preferential activation for schematic faces as well as for simple cartoon characters such as Mickey Mouse.[13] Because our face-detection system consistently responds to these minimally face-like stimuli, it is plausible that many robots will meet the requirements for triggering face detection, thereby triggering AGENT categorization.

The presence or absence of a face has been shown to affect subjects' judgments about robots, specifically. In a recent paper on attributions of experiences to robots, K. Gray and Wegner (2012) showed participants video clips of a lifelike robot. In one condition, the human-like face was visible; in the other condition, the robot was filmed from behind. In both conditions, the robot moved around, and participants were asked to indicate their level of agreement with statements such as "This robot has the capacity to feel pain" and "This robot has the capacity to feel fear" (K. Gray & Wegner, 2012, p. 126). They found that participants gave significantly higher ratings of the capacity for emotion when the face was visible than when it was not.[14]

3.3 Contingent Interaction: ELIZA Effect

The most familiar form of contingent interaction with machines comes not from robots, but computers. Some of the oldest (and most anecdotal) work remains compelling. In the 1960s, Joseph Weizenbaum created a very simple chatbot, ELIZA. The program has canned responses to certain inputs, and often incorporates pieces of user input into subsequent responses. For instance, if you tell ELIZA, "I'm afraid of heights," the bot responds, "How long have you been afraid of heights?" or "Do you believe it's normal to be afraid of heights?" Given how simple the program is, if one starts quizzing ELIZA, it quickly becomes apparent that the program is extremely limited and stupid. However, so long as the brittleness of ELIZA is not exposed, this incredibly thin slice of contingent interaction leads people to flood attributions of sensitivity, empathy, and intelligence to the program: "The most superficial syntactic tricks convinced some people who interacted with ELIZA that the program actually understood everything they were saying, sympathized with them, even *empathized* with them" (Hofstadter, 1996, p. 158). This inclination to inflate attributions of mental states is now known as "the ELIZA effect."

As far as we know, there has been no systematic social scientific research on reactions to ELIZA. But we ourselves feel the pull of ELIZA: So long as the program's brittleness is not exposed, it is easy to slip into thinking that one is interacting with an intelligent being. Moreover, there has been significant work on the treatment of computers that interact with users. For instance, in some experiments on the ultimatum game, participants play against a computer. Strikingly, people often forego money in order to reject "unfair" offers made by a computer (e.g., Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; van't Wout, Kahn, Sanfey, & Aleman, 2006).

People also show a kind of reciprocity to computers. Fogg and Nass (1997) had participants engage in a task in which a computer was supposed to help the participants find information. After this, participants were asked to do another task in which they were in a position to help the computer by providing information on color contrasts. In one condition, participants performed this helping task on the very same computer they used for the first task; in the other condition, participants used a different (but type-identical) computer to perform the second task. They found that participants did significantly more work when they were assigned to the computer that had helped them previously. In yet another study, Nass, Steuer, and Tauber (1994) found that participants exhibit a surprising kind of "politeness" toward computers. When evaluating a person's performance, people are much less critical if the evaluation is given to the person him- or herself. The natural interpretation of this is that we do not want to hurt the person's feelings. As it happens, people do the same thing with computers: If asked to evaluate a computer's performance, people give less critical evaluations when the evaluation is completed on the computer that is being evaluated (as compared to evaluations completed on another computer or by paper and pencil; Nass et al., 1994). One natural interpretation of these results is that there is some implicit concern about hurting the computer's feelings.

3.4 A Fuller Robotic Agent

The previous studies each used only one of the Agency cues. In a recent study with an iCat, the robot exhibited multiple Agency cues, and the purpose was to determine whether participants would resist wiping out the memory and "personality" of a robot. Bartneck, Van Der Hoek, Mubin, and Al Mahmud (2007) programmed an iCat robot to move its face in ways that mimic human expressions. The robot was also programmed to cooperate with a human participant in playing a game of Mastermind against a computer player. The participants were told they would play a collaborative game with the robot in order to develop the robot's personality. They were also told that after the game, they would have to turn off the robot, permanently erasing its memory and personality (Bartneck et al., 2007, p. 219).

The experiment was a 2 × 2 design varying intelligence and agreeableness. For the high intelligence condition, the robot gave smart suggestions for playing the game; for the low intelligence condition, the robot gave weak suggestions. In high agreeableness, the robot was polite, for example, asking if it could make a suggestion; for low agreeableness, the robot was abrupt, for example, insisting on its turn.

After the game, the experimenter communicated by a walkie-talkie, telling the subject, "You can now switch the iCat off." The robot protested, saying things such as "You are not really going to switch me off, are you?" The participants often engaged in conversation with the robot, and showed significant hesitation in turning off the robot. In addition, both high agreeableness and high intelligence increased the hesitation significantly. Indeed, when the robot exhibited high intelligence and high agreeableness, participants took almost 3 times as long to turn off the robot as when the robot was low in intelligence and agreeableness (34.5 s vs. 11.8 s; Bartneck et al., 2007, p. 221).

Bartneck and Hu (2008, p. 420) conducted an even more dramatic experiment, in which subjects were instructed to interact with and then "kill" a robot, by smashing it with a hammer. The robot was a "Crawling Microbug," equipped with light sensors and programmed to move toward the subject's flashlight either clumsily (in the "stupid" condition) or efficiently (in the "smart" condition). The robot was considered "dead" when it stopped moving and its lights stopped flashing; then, the "Number of Hits," "Number of Pieces," and "Level of Destruction" were recorded. These quantitative data are not terribly informative, because many orthogonal factors contribute to the number of hits and number of pieces a robot is smashed into (e.g., arm strength, accuracy, and so on). But subjects' postexperiment remarks are revealing:

> Several participants commented that: "I didn't like to kill the poor boy," "The robot is innocent," "I didn't know I'd have to destroy it after the test. I like it, although its actions are not perfect," and "This is inhumane!" (Bartneck & Hu, 2008, p. 426)

Remarks such as these suggest that although subjects succeeded in destroying the robot, on some level, they conceived of the Microbug as something that can feel pain. One way to understand what's going on here is that subjects have a low-road inclination to think of the Microbug as a pain feeler, yet use high-road reasoning to either override the pain attribution or rationalize causing harm to a feeling thing.

All of the studies we've reviewed in this section have significant limitations. But as a whole, they make a strong case that we are naturally inclined to attribute a wide range of mental states, including states that are phenomenal, nonphenomenal, valenced, unvalenced, and sometimes unrelated to the robot's functional purpose. People often resist attributing mental states to robots, but this, we think, is driven by a high-road process that invokes the culturally prevalent platitude that robots do not have minds. When actually interacting with robots, on the other hand, it seems to be natural for people to attribute mental states to the machines.

Conclusion

When we probe people for their explicit judgments about whether robots have mental states, responses are influenced by a wide variety of factors. The apparent function of the robot, the nature of the question (forced choice vs. not), and platitudes about robots may all contribute to producing reasoned judgments about the states of robots. But there is also a more fundamental tendency to treat robots as fully minded. In ecologically valid settings, this low-level tendency tends to manifest in the form of automatic, unreasoned attribution of a wide range of mental states to robots. We can best understand the overall pattern of folk attributions by distinguishing the roles of high-road and low-road processing and by separately examining their respective contributions to mental state attribution.

Many questions about attributions to robots remain unanswered. In particular, little has been said about how people come to attribute specific kinds of mental states to robots. We have argued that neglecting the distinction between low-road and high-road processing leaves us with an incomplete picture, but we have not offered an alternative positive account that explains why people ascribe particular kinds of states to robots under various circumstances. Similarly, other accounts on offer attempt to explain why people *resist* attributing certain kinds of mental states to robots in vignette studies. But what specific factors drive us to attribute to robots the particular mental states that we *do* attribute? This question is ripe for future research, which should be pursued in light of our low-level capacity for mind-detection, in addition to our high-level considered judgments about robots.

Notes

1. Sytsma and Machery (2010) explicitly interdefine "valence" and "hedonic value": "Throughout we will use the term "valence" as follows: mental states have a valence if and only if they have a hedonic value for the subject. That is, mental states have a valence if and only if they are pleasurable (they then have a positive valence) or disagreeable (they then have a negative valence)" (p. 300).
2. Although isoamyl acetate is an ester that has a strong banana-like odor (for humans), it is reasonable to suppose that subjects in the study did not know this fact about the unfamiliar compound.
3. Buckwalter and Phelan (2013) argue for the conclusion that the folk are what philosophers would call "analytic functionalists," and use the results from the experiment presently under discussion in an attempt to support this view. But this attempt is problematic, for at least two reasons. According to "analytic" versions of functionalism (e.g., Armstrong, 1968; Lewis, 1972), causal roles supply the reference-fixing conditions for mental states, and the relevant causal roles are determined by analysis of our ordinary concepts of mental states (i.e., the "platitudes" about the mental states). So, on traditional analytic functionalist views, the relevant notion of "function" is a causal one. But because Buckwalter and Phelan's Jimmy-experiment exploits a purposive or teleological notion of function rather than a causal one, it is hard to see how the data could support the thesis that ordinary people "are analytic functionalists." A second and more serious problem is that traditional analytic functionalism is a complex view about the *relationship* between ordinary mental state concepts and the metaphysical nature of mental states. It is doubtful whether ordinary people are committed (even implicitly) to a thesis like this, irrespective of their judgments about particular cases. While examining people's intuitions may well reveal features of their mental state concepts (and perhaps which features are platitudinous), simply uncovering how people wield such concepts does not establish that they implicitly hold the more nuanced philosophical view.
4. The first kind of consideration focuses on the degree to which an entity is capable of engaging in goal-directed behavior, the second kind on "states that allow an entity to be concerned with how things go for her" (Huebner, 2010, 151). Huebner associates the first with Dennett's "intentional stance" and the second with Dennett's "personal stance," but he does not provide much more detail.
5. See Stanovich and West (2000) for a fuller explanation of theories about dual-process cognition.

6. Clearly these cues are not the only things that can trigger the identification of an entity as an AGENT. But these cues have been the focus of the key research that we draw on.
7. Also noteworthy is that a majority of subjects judged that insects actually can feel pain.
8. We changed the color of the moved box from red to green because pilot studies suggested that some participants took a metaphorical interpretation of "saw red" to mean *angry*.
9. Participants are less likely to choose "saw green" than either "detected green" (McNemar's test, $N = 25, p < .05$) or "located the green box" (McNemar's test, $N = 25, p < .001$).
10. Participants are less likely to choose "knows the location of the green box" than either "can detect the location of the green box" (McNemar's test, $N = 27, p < .05$) or "processes the location of the green box" (McNemar's test, $N = 27, p < .05$).
11. Tremoulet and Feldman (2000) had participants judge whether the object was "alive", but attributions of "alive" might well be equivalent to (or closely parallel) attributions of Agency (see Arico et al., 2011, pp. 344–346).
12. The SAM scale is traditionally used as a pictorial means of self-reporting emotions. Saerback and Bartneck (2010) converted it into a pictorial technique for emotion attribution.
13. The use of relatively simple face-cues in robots might be desirable in order to avoid the "uncanny valley" effect (Mori, 1970). When a robot looks and acts nearly exactly like a human, but not exactly like a human, people tend to find its appearance disturbing and revolting. By designing robots with simple and/or exaggerated features (as in a schematic or cartoon face), the robot avoids the uncanny valley effect because it does not approach exact realism.
14. It should be noted that on the whole the attributions of experience were fairly low, as were attributions of agency (e.g., the capacity to plan actions). In all cases, the mean attributions were on the "disagree" end of the spectrum (Gray & Wegner, 2012, p. 127). Here, as in vignette studies, it may be that subjects' attributions reflect some high-road processing, because they had plenty of time to make their judgments.

References

Arico, A., Fiala, B., Goldberg, R., & Nichols, S. (2011). The folk psychology of consciousness. *Mind & Language, 26*, 327–352.

Armstrong, D.M. (1968). *A materialist theory of the mind.* London, England: Routledge.

Bartneck, C., & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies, 9*, 415–433.

Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007). "Daisy, daisy, give me your answer do!" switching off a robot. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction, Washington DC* (pp. 217–222). New York, NY: ACM.

Bradley, M., & Lang, P. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry, 25*, 49–59.

Buckwalter, W., & Phelan, M. (2013). Function and feeling machines: A defense of the philosophical conception of subjective experience. *Philosophical Studies, 166*, 349–361.

Fiala, B., Arico, A., & Nichols, S. (2012). On the psychological origins of dualism: Dual-process cognition and the explanatory gap. In E. Slingerland & M. Collard (Eds.), *Creating consilience: Issues and case studies in the integration of the sciences and humanities* (pp. 88–109). Oxford, England: Oxford University Press.

Fogg, B.J., & Nass, C. (1997). How users reciprocate to computers: an experiment that demonstrates behavior change. In *Extended Abstracts of the CHI97 Conference of the ACM/SIGCHI* (pp. 331–332). New York, NY: ACM.

Gray, H., Gray, K., & Wegner, D. (2007). Dimensions of mind perception. *Science, 315*, 619.

Gray, K., & Wegner, D. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*, 125–130.

Guglielmo, S., & Malle, B.F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin, 36*, 1635–1647.

Hadjikhani, N., Kveraga, K., Paulami, N., & Ahlfors, S. (2009). Early (N170) activation of face-specific cortex by face-like objects. *Neuroreport, 20*, 403–407.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243–259.

Hofstadter, D. (1996). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* New York, NY: Basic Books.

Huebner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the Cognitive Sciences, 9*, 133–155.

Johnson, S. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London B, 358*, 549–559.

Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? Features that elicit gaze-following in 12-month-olds. *Developmental Science, 1*, 233–238.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*, 249–258.

Mori, M. (1970). *Bukimi no tani* [The uncanny valley]. *Energy, 7*, 33–35.

Nass, C., Steuer, J., & Tauber, E.R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on human factors in computing systems: Celebrating interdependence* (pp. 72–78). Boston, MA: ACM.

O'Craven, K.M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience, 12*, 1013–1023.

Saerbeck, M., & Bartneck, C. (2010). Perception of affect elicited by robot motion. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 53–60). Piscataway, NJ: IEEE Press.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., & Cohen, J.D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*, 1755–1758.

Stanovich, K., & West, R. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645–726.

Sytsma J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies, 151*, 299–327.

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response properties of the human fusiform face area. *Cognitive Neuropsychology, 17*, 257–279.

Tremoulet, P.D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception, 29*, 943–952.

van't Wout, M., Kahn, R.S., Sanfey, A.G., & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental Brain Research, 169*, 564–568.

Wheatley, T., Milleville, S.C., & Martin, A. (2007). Understanding animate agents: distinct roles for the social network and mirror system. *Psychological Science, 18*, 469–474.