If Asimo Thinks, Does Roomba Feel? The Legal Implications of Attributing Agency to Technology

Christopher Brett Jaeger, Daniel T. Levin
Vanderbilt University

Just as our interactions with other people are shaped by our concepts about their beliefs, desires, and goals (i.e., “theory of mind”), our interactions with intelligent technologies such as robots are shaped by our concepts about their internal operations. Multiple studies have demonstrated that people attribute anthropomorphic features to technological agents in certain contexts, but researchers remain divided on how these attributions arise: What default assumptions do people make about the internal operations of intelligent technology, and what events or additional information cause us to alter those default assumptions? This article explores these open questions and some of their implications for law and policy. First, we review psychological research exploring people’s attributions of agency, with particular focus on attributions to technological entities. Next, we define and describe one popular account of this research—a “promiscuous agency” account that assumes a reflexive tendency to broadly attribute humanlike properties to technological agents. We then summarize mounting evidence that people are often more cautious in attributing human properties than the promiscuous agency account suggests. We seek to integrate the mounting evidence for a “selective agency” account with the promiscuous agency account through the transition model of agency. Finally, we explore how selective agency, promiscuous agency, and the transition model relate to a sample of robotics law and policy issues. We address, in turn, issues related to Fourth Amendment protection, copyright law, statutory and regulatory interpretation, and negligence litigation, identifying specific implications of the transition model of agency for each issue.

Keywords: agency, concepts, human-robot interaction, theory of mind, law, negligence, copyright, Fourth Amendment, privacy, policy, driverless cars, robotics

1. Introduction

Intelligent machines now perform many tasks that were historically performed exclusively by humans. Computers automatically generate articles that we read in the newspaper. Prototypes of Google’s driverless car are navigating our roads. In some places, humanoid robots act as caregivers for the elderly (for review see Broadbent, Stafford, & MacDonald, 2009; Richards & Smart, 2016). As we interact with this growing variety of intelligent machines, the way we conceptualize and interpret their operations will be increasingly important—both on an individual level, as we do our best to navigate our daily encounters with technology, and on a societal level, as we develop law and policy to govern intelligent machines and our interactions with them.

For example, imagine that, in the future, a hospital hires you to manage its internal pharmacy. Most of the work in the pharmacy is completed by a team of robots. A central robot housed within the pharmacy monitors and fills prescriptions for patients. Throughout the day, this central robot gives the filled prescriptions to mobile pill-dispensing robots that deliver the pills to patients’ room. Your job is to oversee this automated process and make sure everything goes smoothly (and you understand that there have been no problems since the system was implemented several years ago). On your third day of work, one of the pill-dispensing robots is picking up medicine for a cancer patient named Peter Paulson when the central
robot gives it some pills. Upon receiving the pills, the pill-dispensing robot says, audibly, “What is this medicine?” Your mind rapidly fills with questions. Is the statement directed at you? Is it directed at the central robot? Do robots speak to each other in English? Does the pill-dispensing robot even “know” that the central robot is a robot? Most pressing, are you witnessing a simple confirmation sequence that is supposed to happen intermittently, or is the pill-dispensing robot concerned about something? As you ponder what you should do, the central robot states the name of the medicine that Peter Paulson is supposed to receive. Seemingly satisfied, the pill-dispensing robot exits the pharmacy, travels to Peter’s room, and delivers the pills, which Peter takes. Unfortunately, they are the wrong pills, and they cause Peter severe injury. Was this your fault? The fault of one of the robots? The robots’ manufacturers? The hospital? Who would be liable if Peter decided to pursue a tort claim?

Our answers to questions like these are informed by the specific attributions we make concerning the robots’ operations. Many of these attributions are shaped by our concepts of other people’s beliefs, desires, and goals (our “theory of mind”), a topic that psychologists have studied extensively. This literature provides a good start at understanding how people might think about intelligent technological agents, but the answers to many important questions (and, in some cases, the important questions themselves) are only beginning to take shape. This article focuses specifically on three interrelated questions:

(i) What default assumptions do we make about what goes on inside a technological agent?
(ii) What events, information, or processes cause us to alter those default assumptions?
(iii) When we change our default assumptions, how broad are the changes?

A body of literature has developed around the first two questions, while the third question remains relatively unexplored.

In reviewing the existing scientific literature surrounding the first two questions, we identify a popular account of the findings that we refer to as the “promiscuous agency” account. On this account, people tend to automatically attribute humanlike qualities to technological agents and only pare back these broad attributions with motivation and deeper consideration (e.g., Epley, Waytz, & Cacioppo, 2007). But more recent research contrasts with the promiscuous agency hypothesis (e.g., Levin, Adams, Saylor, & Biswas, 2013). We describe these non-conforming data and then explain how they constitute a “selective agency” pattern of findings, whereby people are inclined to limit attributions of agency to nonhumans and only make such attributions after deeper consideration. After describing these findings, we examine the ways they can potentially be reconciled with the “promiscuous agency” account to explain both sets of results. An important means of reconciling these two hypotheses is to propose that a relatively heterogeneous collection of cognitive processes control attributions of agency and that situations vary in how they draw upon these processes. This proposal is formalized in the transition model of agency. We describe how the transition model can account for attributions of agency in a wide variety of settings and further explain how it can help us understand how these attributions might change with experience.

Finally, we turn to the implications of promiscuous agency, selective agency, and the transition model for law and policy. We examine a diverse sample of law and policy issues discussed in the robotics law literature, including issues related to the Fourth Amendment, copyright law, statutory and regulatory interpretation, and negligence litigation. For each issue, we explain how and why attributions of agency are relevant, drawing heavily from the transition model of agency. We identify specific features of transition model that are particularly relevant to each issue and explore how those features might interact with real-world factors ranging from product design to decision-maker expertise. Ultimately, we intend for these specific discussions to illustrate a broader point: a thorough empirical account of attributions of agency can, and should, play an important role in shaping law and policy about intelligent technologies.

---

1 When we use the phrase “attributions of agency” in this article, we are referring to attributions of human-like agency, or anthropomorphic attributions. While there may be a theoretical argument that people could attribute some other form of agency to technological agents, that argument is beyond the scope of this article.

2 The set of issues that we discuss is emphatically not exhaustive. Nevertheless, by addressing a diverse set of legal issues, we aim to illustrate the potentially broad legal significance of anthropomorphic attributions.
2. **Attributions of Agency: Findings and Explanations**

2.1. **Agency Concepts**

Social interaction is a foundational component of the human experience, so it is unsurprising that a wealth of psychological research explores how people think about and behave toward others. Historically, this research has focused on thoughts and behavior toward other people. Yet, in order to successfully navigate our world, we must interact with a multitude of entities that are not people. In these interactions, it is helpful, if not essential, for us to distinguish things that are capable of thinking and engaging in goal-directed behavior from those that are not.

As demonstrated by Woodward (1998), we develop the ability to distinguish goal-driven agents from non-goal-driven objects early in life. In Woodward’s experiments, infants repeatedly observed either a human actor’s hand or an inanimate stick reaching for one of two toys (a bear and a ball) on a stage. After enough repetitions, infants habituated to the scene—meaning their response (measured in looking time) decreased until hitting a minimum. After habituation, the locations of the two toys were swapped for a test trial. On the test trial, the human hand or stick either reached to the same location for a different toy or to a different location for the same toy. Nine-month-old infants looked longer (indicating surprise) when the human hand reached for a different toy than when the human hand reached for the same toy in a different location. Infants who saw the inanimate stick did not show this pattern of response. These results suggest that very young infants both understand the concept of goal-driven behavior and limit it to humans.

This early understanding of goal-driven behavior provides a foundation for the subsequent development of “theory of mind”—a more complete understanding that others have beliefs, desires, and intentions different from oneself, and that these beliefs, desires, and intentions influence behavior. Theory of mind has frequently been studied with false belief tasks. For instance, Wimmer and Perner (1983) told stories to their three- to nine-year-old participants in which one character placed an object in one location in the presence of a second character, then transferred the object to a new location after the second character had clearly left the scene. The children were asked where the second character would look for the object upon his return. The three-year-old participants tended to “flunk” the test, mistakenly predicting that the puppet would look for the object in the second location, while four-year old participants made fewer mistakes. This finding, replicated many times, is typically interpreted as evidence that the younger children do not understand that the second character in the story has beliefs that are different from their own. Thus, it has been suggested that “theory of mind” entails two distinct types of cognitive processes: early-developing, basic, automatic processes that account simply for whether others are intentional agents (as demonstrated in the Woodward experiments), and more sophisticated, non-automatic, relatively resource-intensive processes needed to track the beliefs and perspectives of others more completely—processes that take effort and time even for adults (Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006).

Some researchers, however, question whether the children who fail the false belief task really lack the capacity to represent others’ false beliefs. Kovács, Téglás, and Endress (2010), for example, contend that young children have this capacity, but that it is effectively hidden by other, slower-developing abilities required by the task. Thus, these researchers developed a novel task, in which participants watch a movie of an event involving an animated character, a ball, a table, and an occluder. The basic false-belief variant of the movie begins with the character placing the ball on the table. With the character still present, the ball rolls behind the occluder and stops. The character leaves the scene. Then, in the character’s absence, the ball rolls out from behind the occluder and off the screen. The character returns, and he presumably incorrectly believes that the ball is still behind the occluder. The movie ends with the occluder being lowered to reveal the absence of the ball. (In other variants, the character incorrectly believes that the ball left the screen when it is actually behind the occluder, or the character’s beliefs about the ball’s location are correct because the ball does not move after the character leaves the scene.) The experimenters found that adults’ reaction times in assessing the presence or absence of the ball were influenced by the false beliefs of the character (even though the character’s beliefs were completely irrelevant to the task). The experiment was then adapted to run on seven-month-old infants with the critical measure being looking time (as in the Woodward experiments discussed above). Similar to the adult participants, the infants’ looking times were affected by the character’s false beliefs. Kovács and colleagues concluded that the presence of a social agent “is sufficient to automatically trigger” computation and storage of that agent’s beliefs in adults and seven-month-old infants. While the findings of Kovács et al. have recently come into question due to
potential artifacts in their experimental paradigm (Phillips et al., 2015), there are many similar findings based on different paradigms that also support the idea that children much younger than four years old possess an understanding of false belief (e.g., Onishi & Baillargeon, 2005; Scott, Richman, & Baillargeon, 2015; Yott & Poulin-Dubois, 2012).

The debate over the nature and sophistication of the processes involved in theory of mind is ongoing, and this debate has important implications for another critical question: How broadly are theory of mind processes deployed? As observed by Levin, Adams, et al. (2013), Woodward’s (1998) work with infants seems to suggest that adults would limit deployment of theory of mind processes to other humans, or at least to other living things. But other research suggests deployment of these processes is not so limited. Johnson (2003) found that minimal morphological and behavioral cues can cause infants to treat a novel ambiguous object as an intentional, goal-directed agent. Further, multiple researchers have demonstrated a surprising willingness of both children and adults to attribute agency—or at least attribute certain humanlike properties—to robots (Jipson & Gelman, 2007; Kahn et al., 2012; Kahn, Friedman, Perez-Granados, & Freier, 2006; Melson et al., 2009).

One illustration of children anthropomorphizing a technological agent comes from Kahn et al. (2012). These researchers had 9-, 12-, and 15-year-old participants take part in individual interactions with a humanoid robot named Robovie for 15 minutes. During the interaction, Robovie greeted, conversed with, and ultimately played an “I Spy” game with the participant. Robovie also engaged in a number of traditionally-human behaviors during the interaction, making a prosocial request, apologizing for an error, and asking permission to give the participant a hug. At the end of the session, an experimenter came into the room and instructed Robovie to go to the closet. During the ensuing exchange, Robowie protested going to the closet on several grounds before ultimately being guided to the closet by the arm. The experimenter then engaged the participant in a structured interview about Robovie and his qualities. Among other results, a majority of participants reported that Robovie was intelligent (79%) and has feelings (60%), that Robovie could be their friend (77%), that they would feel a need to comfort Robovie if he said he was sad (81%), and that it was not right for Robovie to have been put in the closet over his objections (54%).

But it is not only children who attribute humanlike qualities to technological entities. Adults, too, will in some cases attribute agency (or at least act as though they attribute agency) to technology. Research by Ellen J. Langer and her colleague indicates that many social interactions are essentially carried on “mindlessly”: driven by “minimal structural cues activating standard behavior scripts” with most semantic information in the exchange never reaching consciousness (for review, see Langer, 1992). Building upon Langer’s research, Nass and Moon (2000) examined the degree to which people interacting with computers “mindlessly” treat them like humans.

Nass and his colleagues essentially replicated social psychology research demonstrating consistent scripts or norms in human-human interactions, but substituting a computer for the human target. In one line of experiments, the researchers demonstrated that people rely on social categories when interacting with computers, even when it makes no sense to place a computer in those categories. For instance, the researchers had participants interact with male-voiced or female-voiced computers and demonstrated that participants gender-stereotyped computers in the same ways they gender-stereotype other humans (Nass, Moon, & Green, 1997). Additional experiments in this vein further demonstrated the influences of racial categorization and in-group/out-group effects (Nass, Fogg, & Moon, 1996; Nass, Isbister, & Lee, 2000).

Another line of experiments demonstrated that people apply social norms like politeness and reciprocity in interacting with computers (Nass, Moon, & Carney, 1999; Fogg & Nass, 1997; Reeves & Nass, 1996). A third line demonstrated that people are attuned to a computer’s “personality,” with the degree of match between the participant’s personality and the computer’s “personality” informing participants’ attributions of intelligence to the computer, disposition toward the computer, and even willingness to use the computer to purchase items (Moon, 1998; Moon & Nass, 1996; Nass, Moon, Fogg, Reeves, & Dryer, 1995). Interestingly, Nass et al. noted that none of their participants ever reported believing that a computer should be understood or treated as a human, which they note stands in “stark contrast to people’s actual behavior” in the experiments (Nass & Moon, 2000).

---

3 Robovie, while capable of autonomous functioning, was controlled wirelessly by experimenters during the study.
2.2. The Promiscuous Agency Account

After analyzing a wide variety of research on attributions of agency—including many of the articles discussed above—Epley, Waytz, and Cacioppo (2007) sought to provide a systematic account of the human tendency to anthropomorphize. They observe, quite correctly, that anthropomorphism is a “process of inference about unobservable characteristics of a nonhuman agent”—a special case of induction. They then proceed to describe their hypothesis of how the different steps of the general inductive process apply to reasoning about a nonhuman agent.

First, to draw an inference, one must first activate knowledge about a known, readily-accessible base for induction. In the case of anthropomorphic induction, this base is knowledge about humans generally, or oneself specifically. Critically, per Epley et al., people automatically use this base for induction about the qualities of nonhuman agents. For example, if you are interacting with a robot, and you happen to have substantial knowledge about how robots operate (e.g., if you happen to be a robotics engineer), you are more likely to correct your bias toward anthropomorphism—and thus less likely to anthropomorphize the robot—than the average person, at least if you have reason to access your knowledge about robots. The second factor, effectance, refers to motivation to feel effective in, and reduce uncertainty about, one’s environment and in one’s interactions with other agents in it. For example, if you encounter a machine that is particularly unpredictable or particularly threatening, you will be less likely to correct your bias toward anthropomorphism—and thus more likely to anthropomorphize the machine—than if the machine were predictable or non-threatening. The third factor, sociality, refers to mankind’s need for social connections. The idea here is that, if you become increasingly lonely, you will be increasingly unlikely to correct your bias toward anthropomorphism—and thus more likely to anthropomorphize—nonhuman entities.

Epley et al. go on to support their promiscuous agency account with examples of existing research that it fits nicely. For instance, Epley and colleagues assert that, based on the “elicited agent knowledge” factor, people are more likely to anthropomorphize technological agents that physically resemble humans. They then identify findings that support this assertion in the literature (Bergou et al., 2000; D’Salvo, Gemperle, Forlizzi, & Kiesler, 2002). Similarly, Epley and colleagues contend that their predictions concerning the “effectance” factor were supported by prior research demonstrating that participants became more willing to anthropomorphize computers the more frequently those computers malfunctioned (Morewedge, 2007). With respect to loneliness, Epley, Akalis, Ways, and Cacioppo (2008) found that participants experiencing social disconnection were more likely to report beliefs in the supernatural and to anthropomorphize pets; however, other experimenters were unable to replicate this result (Sandstrom & Dunn, 2015).

It is important to note that Epley et al. are not the only researchers that have taken a “promiscuous agency” view of agent reasoning. Barrett and Lanman (2008), for instance, connect religious beliefs to (over-)attributions of agency and theory of mind. Specifically, they propose that human beings are equipped with a cognitive/perceptual subsystem, referred to as the “Hyperactive Agency Detection Device” (HADD), which identifies agents based on simple and minimal perceptual cues. They further propose (based on evolutionary logic and experimental findings) that this subsystem is “hypersensitive,” and that once it identifies an agent (whether real or illusory), the mind makes a plethora of inferences about the intentions of the agent. Barrett and Lanham echo Epley, Waytz, and Cacioppo’s promiscuous agency account in contending that people are initially inclined to (over-)attribute agency to ambiguous stimuli and that subsequent processes may scale back these initial attributions.

In sum, a robust set of findings demonstrates that people sometimes make anthropomorphic attributions to nonhuman entities, and at least two groups of researchers have posited similar systematic account of how those attributions come to be. We refer to that account as the “promiscuous agency account” because it posits that, when confronted with an unknown or unpredictable stimulus, people default to anthropomorphizing it. The initial inclination toward anthropomorphism is only scaled back through deeper consideration.
2.3. The Limits of Promiscuous Agency and the Selective Agency Account

The promiscuous agency account explains some of the empirical findings about anthropomorphic attributions, but it does not explain all of them. The first component of the promiscuous agency account—the proposed automatic bias toward anthropomorphizing the unknown—is difficult to reconcile with other findings. First, developmental data demonstrates that children sometimes draw sharp distinctions between living and nonliving agents (Gutheil, Vera, & Keil, 1998). Second, more recent research examining the extent to which adults distinguish between the internal (mental) processes and representations of humans, computers, and robots, reveals similar sharp distinctions. This section discusses some of this more recent research.

In one set of experiments, Levin, Killingsworth, Saylor, Gordon, and Kawamura (2013) asked adult participants to make predictions about the behavior of a computer, a human, and a robot in two types of scenarios. The first type was an object-versus-location scenario modeled on the Woodward (1998) infant paradigm: Participants viewed two objects (a toy duck and a toy truck) on a 3x3 grid and were told that each of three agents (a human, a computer, and a robot) picked up the duck from Location A1 in the first two trials of a three-trial sequence. For the third trial, the locations of the duck and truck are switched, and participants are asked which object each agent would “pick up” on the third trial: the truck in Location A1 (as one would expect if the agent acts in a rote, location-driven manner), or the duck in its new location (as one would expect if the agent acts in an intentional, goal-directed manner, see Fig. 1 for illustration). The questionnaire included a second type of scenario focused on categorization, asking participants to predict whether each of the three agents would divide a set of six objects based on perceptual surface features (dark rectangular objects vs. small colorful objects, as one would expect if the agent did not consider function) or on taxonomic features (food vs. office supplies, as one would expect of an intentional agent applying knowledge of the object’s function, see Bloom, 1997).

Across a series of experiments, participants consistently made significantly fewer “intentional” predictions—predictions that the agent would pick up the same object in a different location in the first type of scenario, and predictions that the agent would group based on a taxonomic category in the second type of scenario—for the computer than for the human (Levin, Killingsworth, et al., 2013). Participants were also hesitant to make “intentional” predictions about the robot. In the first experiment, even though the robot was anthropomorphized in the scenario (given the name “OSCAR” and described as having goals), participants did not make any more “intentional” predictions for the robot than they did for the computer.
In the second experiment, participants viewed a video of “OSCAR” walking, running, and stopping to let humans pass before responding to the scenarios, but still did not make more “intentional” predictions for the robot than the computer. In the third and fourth experiments, participants were shown a video of “OSCAR” looking at ten pairs of objects and instructed to remember which object in each pair “OSCAR” preferred. Only with this strong manipulation did participants make more “intentional” behavioral predictions for the robot than the computer. Taken together, these studies demonstrate “not only a strong contrast in subjects’ intuitions about the cognitive processes inherent in different entities,” but also “that simple anthropomorphism is not sufficient to overcome this difference, at least for the kind of explicit behavioral predictions” tested in the experiments (Levin, Killingsworth, & Saylor, 2008).

Reinforcing this line of experiments, Levin, Killingsworth, and Saylor (2008) asked participants to make similar behavioral predictions about a human, a computer, and a robot, but this time the experimenters varied whether the human, computer, and robot were from the present day or from 100 years in the future. As in the earlier experiments, participants’ predictions for the computer and the robot differed from their predictions for humans. The fact that these differences persisted even when the technological agents were described as being from the future suggests that people do not believe the agent’s limits are simply a product of the limits of contemporary technology. Further, in both sets of experiments, participants were asked to rate how intelligent they believe computers are, and there was no relationship between participants’ ratings of computer intelligence and their behavioral predictions (Levin, Killingsworth, et al., 2013; Levin, Killingsworth, & Saylor, 2008).

Levin, Harriott, Paul, Zhang, and Adams (2013) again replicated the finding that participants predict more intentional behavior for humans than robots and computers using the same scenarios. Interestingly, in this study, participants completed a realistic disaster-response medical triage scenario with either a human-partner or a robot-partner prior to responding to the behavior prediction scenarios. In both the human-partner and robot-partner conditions, participants again strongly distinguished between the human and the computer as well as between the human and the robot.

An additional study using the behavioral prediction scenarios demonstrates that seventh grade students draw distinctions very similar to adults’ (Hymel, Levin, Barrett, Saylor, & Biswas, 2011). Participants were assigned either to an experimental condition, in which they used an agent-based, online tutoring system called “Betty’s Brain” to complete lessons, or a control condition, in which they used traditional classroom assignments. After completing their learning sessions and post-tests about the material, participants responded to three behavioral prediction scenarios with respect to a human, a computer, and Betty. Like adults, the seventh graders made more intentional predictions for the human than for either Betty or the computer and did not distinguish Betty from the computer.⁴

Another line of research showed a similar pattern of results even though participants were engaged in a very different task—deciding who is at fault when intelligent machines cause harm (Jaeger & Levin, in prep). Participants read vignettes describing an incident in which a self-driving car, a drone, or a robot injured someone while functioning autonomously. Some participants’ vignettes featured highly anthropomorphic descriptions of the relevant machines, while other participants’ vignettes included non-anthropomorphic, mechanical descriptions. Subsequently, when asked about the machine at issue, participants in both conditions were highly selective with their attributions of agency. That is, participants were, across the board, very close to floor on the attribution scale developed by Waytz, Cacioppo, and Epley (2014). When the scale was adapted to ask participants about a wider variety of properties, differences between conditions emerged, but—importantly—those differences were restricted to only some types of properties (e.g., the ability to make decisions) and not others (e.g., the ability to feel emotions).

Together, these findings suggest that people are inclined to draw sharp, fundamental distinctions between the internal operations of humans and the internal operations of technologies. Nevertheless, one could conceivably argue that these findings are not incompatible with the promiscuous agency account. While the promiscuous agency account suggests that people are biased to anthropomorphize technological agents, it also allows that the bias can be corrected with sufficient effort. Thus, perhaps participants in these experiments initially anthropomorphized the computers and robots at issue, but something in the experiments led them to pare back their anthropomorphic attributions.

Levin, Saylor, and Lynn (2012) addressed this possibility with two additional experiments focusing on

---

⁴ Interestingly, though, participants who demonstrated an advanced understanding of human agents in the behavioral prediction scenarios learned more effectively from the agent-based Betty’s Brain system.
the relationship between participants’ attributions of agency and their reaction time. If, per the promiscuous agency account, people anthropomorphize first and “correct” later, one would expect the fastest-responding participants to exhibit the *weakest* contrast between humans and machines (as slower-responding participants have more time to “correct” their default bias to anthropomorphize). But, in experiment one, the researchers found the inverse: The fastest-responding participants exhibited the *strongest* contrast between humans and machines. Then, in experiment two, the researchers instructed half of the participants to answer quickly based on first instinct, and instructed the other half to delay their responses after deep consideration. Consistent with the first experiment, the participants in the fast-response condition exhibited stronger contrast between humans and machines than participants in the slow-response condition.

Also difficult to reconcile with the promiscuous agency account is the finding that participants often limit their attribution of agency across properties as well as across agents. Although this aspect of agency attribution is rarely explored directly, it is at least implicit in research (e.g., Grey, Grey, & Wegner, 2007) exploring dimensions of agency, which suggests that instead of attributing some level of overall agency to a given agent, people make independent attributions on different dimensions, such as having agent-like experience (for example, or fear, pain, and desire) and having agent-like cognitive processes (such as self-control, memory, and thought). Accordingly, people might selectively attribute the humanlike experience of desire to an agent while refusing to attribute humanlike memory processes to that same agent.

Recent data documenting middle-school students’ attributions to teachable software agents is consistent with this possibility (Jaeger, Hymel, Levin, Biswas, Barrett, Paul, & Kinnebrew, in prep). After interacting with the teachable agent "Betty" across several science lessons, students were asked a series of questions about the degree to which Betty (versus a generic computer) could be characterized as having a range of mental properties. The students were more willing to judge that Betty was more able to feel, and in some cases to think and to remember, than a computer, but students were never willing to attribute increased ability to have desires to Betty. In addition, students attributed much less intelligence and knowledge to Betty than to the computer. This latter finding may have occurred because the nature of the agent required that it be ignorant of the facts that the students were charged with teaching it, but it makes clear that the students strongly limited their attributions capabilities across properties, and they did so following an even more specific pattern than implied by the Grey, Grey, and Wenger (2007) model.

Other observations converge to reveal selective attributions across properties. For example, as reviewed above, Levin, Killingsworth, et al. (2013) observed that repeatedly viewing a robot look at one of two objects led participants to attribute more agency to the robot. When they assessed which behavioral prediction scenarios revealed this change, they found that only the Woodward-based object vs. location scenario changed (and it changed strongly), while other scenarios, like the feature vs. category scenario, did not change at all. They argued that the object vs. location scenario changed because it revealed object-related goals in a situation where an agent chose from one of two objects in a situation that was very similar to the observations that participants made when a robot was looking at and responding to one of two objects in the training video. The other scenarios assessed other aspects of intentionality (e.g., having semantic knowledge-structured categories as opposed to perceptual feature-based categories) that might logically connect with having object-related goals, but clearly they are not the same thing, and participants were unwilling to make attributions across these properties. In addition, as mentioned above, Jaeger & Levin (in prep) found a similar pattern: When participants read anthropomorphic descriptions of intelligent machines, they made more attributions of agency to those machines for some capacities but not for others.

Taken together with the previous findings, this pattern of results poses a substantial problem for the promiscuous agency account as a stand-alone explanation of anthropomorphic attributions, both because participants make immediate selective attributions across agents and because they make selective attributions across properties. In fact, this pattern of findings supports precisely the opposite of the promiscuous agency account—what we call the “selective agency account.” The selective agency account posits that people’s default or baseline assumption in dealing with a technological agent is to sharply distinguish the agent from humans, making specific anthropomorphic attributions only after deeper consideration.

While there appears to be a degree of tension between the promiscuous agency account and the selective agency account, we submit that the apparent tension is resolvable—the two accounts do not have to be direct competitors. Neither account needs to provide a universal explanation of attributions of agency. Rather, it may be that each account accurately describes the process of attribution in distinct contexts, and the two can be reconciled within the framework of an overarching model. Nass and Moon (2000) reported that their participants consistently anthropomorphized computers during human-computer interactions but
also stated that “all computer users know a fundamental truth: The computer is not a person and does not warrant human treatment or attribution.” This contrast is not contradictory; different questions and contexts may elicit different attribution processes (with different defaults).

2.4. Reconciling Promiscuous and Selective Agency: The Transition Model

As reviewed in the previous section, there appears to be good evidence that people often limit their attributions of agency across agents, across the features or abilities attributed to a given agent, and across situations. This selective attribution pattern clearly characterizes thinking in many situations, but it is equally clear that other situations are better characterized by broader, more promiscuous attributions of agency. This has led researchers to conclude that people understand agency by drawing upon a collection of contrasting reasoning processes (see, e.g., Kiesler, Powers, Fussell, & Torrey, 2008)—processes that our review classifies as either selective or promiscuous. But while we agree that contrasting reasoning processes are at work, relatively little has been done to specify the relationships among them. We have therefore been developing the transition model of agency attribution to combine a heterogeneous set of processes with a set of principled distinctions that can help generate predictions about when people will or will not anthropomorphize an agent (Levin, Adams, Saylor, & Biswas, 2013). An important goal of the model is to explain not only the conclusions people may draw about agents, but also how those conclusions and the concepts that generate them might change.

A key empirical observation that the transition model is designed to accommodate is that people sometimes default to promiscuous inductions about agency and then pare back upon further consideration, but they also sometimes start with a selective attributions and only broaden their attributions after further consideration. This dual pattern of findings strongly suggests the need both to consider a heterogeneous set of initial assumptions about agents and a heterogeneous set of second-line processes that reflect deeper thinking. In addition, these basic empirical patterns make clear the need to consider factors that lead people to transition from their initial assumptions to second-line processes.

Fig. 2 depicts the transition model and shows how it distinguishes between first-line processes that people use to structure their initial considerations about an agent and second-line processes that they use when thinking more deeply about an agent. Accordingly, the major distinction in the model is not between kinds of processes (such as implicit versus explicit reasoning, or shallow versus elaborative processing) so much as a distinction between different circumstances in which different processes are deployed (e.g., perceptual identification and reasoning leading to an initial conclusion versus subsequent perceptual exploration and reasoning that serves to revise the initial conclusion).

The Transition Model for Agency Cognition

First-line Processing
- Perceptual Detection and Classification
- Basic Agency/TOM and Reference
- Social-Cognitive Defaults
- Local Givens

Cognitive Dissonance
Availability/Accessibility of Second-line Concepts

Second-line Conceptualization
- Appeal to Deep Principles
- Modification of Classification, Entailments, Domain Theory
- Reconceptualization of Default

Figure 2. The Transition Model for Agency Cognition.
First-line processes include perceptual detection and classification processes that can be seen in infants, who respond to differences between living things and nonliving things (Van de Walle, 1999) and then go on to distinguish between actions (e.g., talking vs. touching) that are appropriate for living agents and nonliving things (Molina, Van de Walle, Condry, & Spelke, 2004) and to understand the difference between goal-directed and non-goal-directed actions (Csibra, 2008). Similar processes are implicated in adults’ detection of simple goal-organized movements (Gao, Newman, & Scholl, 2009). In human-robot interaction, perceptual detection and classification processes can be observed when, for example, more agency is attributed to a robot when its movements desynchronize from rhythmic music (Avrunin, Hart, Douglas, & Scaccellati, 2011), or when robots induce increased rapport by matching human gaze patterns (Mutlu, Kanda, Forlizzi, Hodgins, & Ishiguro, 2012).

First-line processes are not limited to perception. It is also useful to consider basic agency/theory of mind and reference—default cognitions that structure basic interactions. For example, young infants follow gaze and by the end of their first year conclude that people represent the things they look at (Woodward, 2003). Similar conclusions support basic social interactions in adults, and these include basic belief tracking (e.g., representation of true beliefs) and understandings of shared attention. Much of the theory of mind literature is predicated on a distinction between such basic cognitive processes and deeper, more effortful cognitive processes involving false beliefs (for review see Baron-Cohen, 1995; Leslie, Friedman, & German, 2004; Samson, Apperly, Braithwaite, Andrews, & Scott, 2010). In addition, behaviors such as those observed by Nass and Moon reveal social-cognitive defaults that reflect widely-studied responses to social scripts. Thus, certain settings and types of interactions may facilitate first-line attributions of agency (e.g., carrying on a conversation about politics with a computer will likely invite more first-line attributions of agency than would typing a paper in Word on the same computer). Finally, we have hypothesized that people often encounter technological agents with limited forms of intelligence and must rely on specific information about the agent’s capabilities. For example, in one study, users guided a robot around an environment by using a Kinect controller to point to locations on the floor (Raza Abdi, Williams, & Johnston, 2013). This explicitly-specified means of interaction is relevant to the robot in that specific study, and users may draw some local conclusions about the robot but likely have little basis for broader conclusions. So, users might conclude that they have to use the Kinect controller because the robot cannot identify hands or arms, while being hesitant to conclude that because the robot must be guided by the Kinect controller, it cannot understand goals more generally. These local givens are often useful bits of information about the capabilities of an agent that users assume are relevant only for a given agent or interaction.

These first-line processes share several properties. They tend to be either specific or shallow in their application. Social-cognitive defaults, for example, represent a fairly broad generalization from the property of interactivity to a reciprocity behavior designed to promote social cohesion (see, e.g., Langer, 1992), but it is also relatively shallow cognitively because it represents a default and inflexible link between one behavior and another without deeper consideration of the entailments of such a broad generalization. As such, these nominally broad generalizations can leave much of the participants’ understanding of the agent unaffected. Similarly, some forms of first-line cognition can be relatively deep in that they may require explicit thought and learning, yet broad generalization is not likely. For example, learning local givens may require considerable effort, but because this effort is viewed as task-specific, it will result in little generalization. Many of these characteristics of first-line processes are shared with related theories of reasoning that also hypothesize a contrast between initial, relatively simple processes and deeper ones (for review see Evans, 2008), although the transition model includes more room for relatively labored first-line reasoning in situations where difficult interaction rules must be learned.

In contrast to first-line processes, second-line processes afford reworking of defaults, deeper reasoning, broader induction, and most important, conceptual change. Second-line processes include appeals to deep principles such as abstract formal relationships that hold across many situations. In the problem-solving literature, this can be seen as processing the schematic structure of a problem as opposed to its surface structure (Gick & Holyoak, 1983). Another second-line process reflects the collections of ways in which a working-memory representation of an agent can be modified. These modifications can involve changes to the classification of an agent (e.g., a technological agent that one initially thought was a computer is reclassified as a robot) or changes to the entailments of the classification (e.g., because the agent is a robot and not a computer, it holds X belief about its environment rather than Y belief). This second type of change includes the effortful, executive-driven selection processes that, according to Leslie, Friedman, and
German (2004), are needed to “pass” false belief tasks—processes that reconfigure one’s default assumptions that another agent’s beliefs, desires, and goals are the same as one’s own. It is interesting to note that these effortful reconfigurations are just the sort of process that would benefit from activating known event schemas, and that in the absence of existing knowledge, particularly difficult reconfigurations may provide the motivation to develop new event-organizing concepts.

Thus, second-line conceptualization also includes potential changes to defaults or possibly the creation of new concepts that may serve as future defaults. For example, if one observes a robot that seems able to understand natural categories (e.g., going beyond merely being able to identify a chair based on its features to being able to truly understand chairs based on their human-centered function), but does nothing without being asked, one could develop a hypothesis that natural category knowledge should be present in many robots. Alternatively, one could develop a new category for entities that have real knowledge but lack goals. This would be consistent with recent proposals that robots present a substantial challenge to ontological categories (Kahn et al., 2011; Kahn, Friedman, Perez-Granados, & Freier, 2006). Finally, individuals who interact with robots could go so far as to reconceptualize their default assumptions, producing feedback from second-line processes to first-line processes. This could occur either as the result of automated learning or more strategic planning to think about agents in a new way.

A key feature of the transition model is that it attempts to enumerate processes that induce transitions between first-line and second-line processes. Based on previous research and experiments testing the transition model, we have included two such factors in the model: the availability of deep and/or broad concepts, and cognitive dissonance.

First, we propose that availability of deep and/or broad concepts may induce transitions. For example, in the case of Guthell, Vera, and Keil (1998) children made broad inductions about unfamiliar properties of known entities, but only when they were reminded about the status of the entities as members of the larger category of living things. Similarly, in adults, large bodies of literature in decision making and problem solving demonstrate that latent knowledge may not affect cognition unless it is made available by hints, priming, or reminders (for review, see Chi & VanLehn, 2012). Much of this literature is focused on the need to deploy relevant knowledge in the service of online reasoning. The converse, however, is also relevant here because the activation of this deeper knowledge also increases the likelihood that it could be changed (for review, see Chinn & Brewer, 1993). The implication of this literature for reasoning about agents is that people possess large amounts of knowledge about agents that may or may not be used in a given situation and so some explanation for what gets activated is necessary.

One such explanation, embedded in the transition model, is cognitive dissonance. Cognitive dissonance can be defined as motivation-producing detection of inconsistencies among new or known bits of information. This well-known concept is borrowed from the attitude change literature (for review see Eliot and Devine, 1994), but our use of it puts less emphasis on ego threat (e.g., the need to avoid looking foolish or morally inconsistent). A similar idea from the learning science literature is that students may experience “cognitive disequilibrium” when they detect inconsistencies, and this leads them to switch from a shallow analysis of a problem to a deeper analysis (Graesser, McNamara, & VanLehn, 2005). In human-robot interaction, cognitive dissonance often results from situations where a robot’s capabilities are surprisingly good or poor. For example, in Levin, Harriott, et al. (2013), participants interacted with a robot or a human-partner in a realistic emergency medical triage scenario and afterward were asked to complete a brief cognitive dissonance questionnaire along with our behavioral prediction measure of agency attributions. In participants who interacted with the robot, we found that high levels of cognitive dissonance predicted lessened attributions of agency to the robot. This suggests that participants experience conflict between their initial beliefs about robots’ capabilities and their experience. This conflict was probably shaped by the specific task, which required participants to verbally describe victim’s symptoms to the robot. This implied that the robot lacked capabilities that might otherwise have been attributed to it. In a more recent study, participants worked with more capable robots and in this case cognitive dissonance was associated with increased attributions of agency.

One relevant set of detectable information conflicts are those between our understanding of an agent’s representation of the world and our understandings of the world’s actual state. The false belief task presents the prototypical version of this conflict. So, one might observe that a robot has experienced the world in a given state, but failed to observe that state change. In such a case, there is nominally a conflict between our

---

5 An event schema is a preconceived idea or mental model of how an event is expected to unfold based on previous experience with similar events.
knowledge that a) the robot thinks that the world is in the initial state, and b) that it is actually in the new state. The need to mediate between these two beliefs and predicate one’s prediction about the robot based on its beliefs as opposed to our belief about the state of the world is the basis for models of theory of mind reasoning that posit the need for higher-order cognition in understanding an agent’s behavior (Leslie, Friedman, & German, 2004). Although this situational constraint might be seen as the kind of cognitive conflict that would produce dissonance and therefore invoke second-line processes, we suspect that the situation lacks a key ingredient necessary to actually invoke dissonance as we have defined it. In fact, many situations where this conflict occurs are routine. People have frequently observed situations where an agent cannot be aware of something because of its recent experience and thus probably have readily available event schemas to guide attention in these situations, and some of these are probably represented as social-cognitive defaults. Therefore, even relatively effortful forms of reasoning about robot representations may induce dissonance only when these basic schemas are disrupted or information is detected that conflicts with them. Accordingly, the transition model assumes that many theory of mind tasks are handled by first-line processes, and that such tasks only invoke second-line processes when substantial expectations are violated in ways that cannot be explained using already-learned schemas.

The transition model makes several important predictions that can be the basis for additional research. Most important, it predicts that first-line processing will rarely produce conceptual change, and that factors leading to second-line conceptualization will induce conceptual change, either directly or indirectly. This is consistent with the literature concerning learning science, in which many researchers propose that cognitive dissonance in a given setting will either directly or indirectly induce conceptual change (for review see Limon, 2001). In addition, because a key feature of second-line conceptualization is generalization, the model provides a principled basis for differentiating contexts where individuals will make only very narrow generalizations across entities, inferred processes, and situations from contexts where those generalizations will be broader. Thus, the transition model accounts for instances in which participants anthropomorphize agents with respect to one characteristic or domain of behavior but do not extend those anthropomorphic attributions to other, seemingly-related domains (e.g., Levin, Killingsworth, et al., 2013).

In sum, the transition model of agency provides a broader framework for understanding when and how people anthropomorphize technological agents than either the promiscuous agency account or the selective agency account can provide standing alone. In fact, the transition model can accommodate both patterns of findings. While the model leaves a number of questions to be answered, we submit that it provides useful scaffolding for future research on anthropomorphic attributions (and their relevance to issues of law and policy).

### 3. Implications for Law and Policy

With an ever-increasing number of intelligent technologies turned loose in (or, perhaps in some cases, on) the world, we, as a society, must begin wrestling with the myriad legal and policy questions these technologies raise. For example, if intelligent technology generates art, whose intellectual property is it? If we have a robot in our home, are we consenting to the robot recording what happens inside? If a self-driving car runs off the road, who is at fault?

A rich and rapidly expanding literature at the intersection of robotics and law identifies and evaluates issues like these among many others (see, e.g., Bridy, 2012; Kaminski, 2015; Calo, 2015; Calo, 2016). Some of the issues are immediate—they concern things that technology already does. For instance, machine-generated art has already been displayed in art museums, though we still have not untangled the intellectual property issues related to such displays (for discussion, see Bridy, 2012). Others issues remain in the future: Should a sufficiently sophisticated robot be treated as a legal person (Solum, 1992; Boyle, 2011) and afforded rights like free speech (see Wu, 2013)? The robotics law literature encompasses both types of issues, aiming to situate emerging technologies within the existing policy landscape while also identifying how that landscape could or should change to accommodate future developments (e.g., Calo, 2016; Richards & Smart, 2016).

Even with the guidance of scholars working at the intersection of law and robotics, the practical undertaking facing our legal system—using and/or updating vast, diverse, and constantly evolving bodies of law to address a vast, diverse, and constantly evolving set of intelligent technologies—presents a tremendous challenge. Further, the challenge will likely be compounded by piecemeal nature of how we will meet it: Inevitably, some aspects of robotics policy will be shaped by legislatures, others by assorted regulatory agencies, and others determined case-by-case by judges and juries (see Calo, 2016; Rachlinksy, 2006).
Critically, whenever and however robotics law issues are resolved, their disposition will depend on someone’s or some group’s beliefs (be they naive assumptions or careful determinations) about the operations of the relevant technology. Thus, the topic at the heart of this article—how people conceptualize the internal operations of intelligent technologies—will be relevant in addressing a wide variety of important issues of law and policy in the years to come.6

In sections 3.1 through 3.4, we discuss a sample of law and policy issues that are likely to be influenced by attributions of agency. These issues include (i) the extent to which the Fourth Amendment protects conduct recorded (only) by an in-home robot, (ii) how to handle copyrights for machine-generated artwork, (iii) statutory and regulatory interpretation of provisions addressing activities that intelligent technologies now, or may soon, engage in (like “driving” or “performing”), and (iv) who faces negligence liability when emergent robot conduct causes injury. We note that all of the issues we discuss have been thoughtfully addressed in the robotics law literature previously. What we aim to add are some thoughts about how the attributions of agency may influence legal and policy decisions on these matters.

3.1. Reasonable expectations of privacy under the Fourth Amendment

It has been nearly a decade since Bill Gates famously called for a robot in every home (Gates, 2007), and while we are not there yet, it seems safe to say that in the coming years, robots will be increasingly prevalent in American households.7 But the benefits of robots in the home in terms of labor reduction necessarily come with tradeoffs in terms of privacy (Kaminski, 2015). In order to perform their intended functions, robots typically sense and record their environment. Further, recorded information is often shared with third parties for processing or cloud storage. This raises two important legal questions: (1) Does granting a robot permission to access a space entail concomitant permission to record information about that space, and (2) does “traditional legal protection of the home as a privileged, private space . . . withstand invasion by digital technology that has permission to be there” (ibid.)? These questions are relevant in various areas of law, from privacy torts to contracts to consumer protection to the First and Fourth Amendments.

Our discussion here will focus on the Fourth Amendment context. Whether a particular law enforcement action runs afoul of the Fourth Amendment’s restriction of unreasonable searches often turns on the target’s “reasonable expectations of privacy.” That is, “an action by law enforcement officers that does not infringe on ‘reasonable expectations of privacy’ is not a ‘search,’ and therefore need not be authorized by a warrant, or be based on probable cause, or be in any other way ‘reasonable’” (Slobogin & Schumacher, 1993). On this basis, it has been held that law enforcement actions such as examining the contents of garbage and searching back yards from flying airplanes do not constitute searches: Individuals have no reasonable expectation of privacy in the things they throw away or things in their yard that are visible from the air (California v. Greenwood, 1988; California v. Ciraolo, 1986). Similarly, under the “third party doctrine,” an individual does not have a reasonable expectation of privacy in information he or she voluntarily reveals to a third party. Thus, statements voluntarily made to undercover agents are not protected by the Fourth Amendment (United States v. White, 1971). The telephone numbers that an individual dials on his or her phone are similarly unprotected, as those numbers are considered voluntarily shared with the telephone company (Smith v. Maryland, 1979).

How, then, will a court treat incriminating video evidence of a criminal defendant that is captured, inside the defendant’s home, by the defendant’s in-home cleaning robot? Does it matter whether the video is stored internally or with a third party cloud service provider? Resolving this issue requires interpretation of the defendant’s reasonable expectations of privacy, which, in turn, depends on the set of attributions that a reasonable person in the defendant’s shoes would make to the robot. This is an area where it seems likely that a court’s decision will be guided by metaphor (see Thibodou & Boroditzsky, 2011). If the in-home robot is viewed anthropomorphically, we would expect the court to find that the criminal defendant had no reasonable expectation of privacy. Conversely, if the robot is viewed more as a device or appliance, we would expect the court to be more likely to find a reasonable expectation of privacy. (The presence of another human in the room destroys a reasonable expectation of privacy; the presence of a vacuum cleaner does not.)

---

6 Metaphors and other mental models can shape policy recommendations and legal outcomes (see Thibodeau & Boroditsky, 2011; Froomkin, 1995).
7 iRobot, the manufacturer of the Roomba, reports having sold more than 14 million home robots worldwide (iRobot, 2016).
Of course, the attributions made to the cleaning robot need not be so rigidly dichotomous. The transition model of agency allows for attributions to be domain-specific—a feature that it seems may matter greatly in this context. For example, even if one attributes to a cleaning robot the capacity to make autonomous decisions about where to clean, one may not attribute the capacities to “notice” (capture) or “remember” (store) information about non-cleaning related events. Thus, anthropomorphistic attributions along some dimensions may be irrelevant, while attributions along other dimensions may be highly relevant, to the defendant’s reasonable expectations.

As to the relevant dimensions, it is possible—if not probable—that a number of seemingly-irrelevant factors (like a robot’s name, form, and intended function) might influence a court’s view of reasonable attributions and therefore its final decision. Video evidence of drugs in a home captured by a Roomba-like cleaning robot that is small and cylindrical may be treated differently than the same information captured by “Geoffrey the Butlerbot,” an extremely-humanoid, social robot that dons traditional butler garb and keeps its owner company in addition to cleaning. With respect to Geoffrey, first-line and second-line processes all may tend to favor attribution—especially if evidence suggests Geoffrey’s functionality is in some ways surprisingly humanlike (e.g., Geoffrey will sometimes fail to spot a stain because it is roughly the same color as the carpet). With respect to first-line processes, an especially humanoid robot would likely invite at least some initial attribution of agency. Even upon deeper consideration, hints to decision-makers invoking the concept of “butler” (via the robot’s name and clothing), together with cognitive dissonance associated with the robot’s humanlike qualities, could further drive anthropomorphistic attribution and consequently decrease the reasonable expectations of privacy of the robot’s owner.

3.2. Copyright protection

As a result of the increasing availability of intelligent technologies, means of artistic production are changing. “[P]eople are enabling computers to produce art and other creative works in new ways, virtually all by themselves” (Bridy, 2012). For instance, artist Harold Cohen spent decades developing a computer program called AARON, which creates original artwork that has been displayed in the Tate Gallery in London, among other places. Similarly, Raymond Kurzweil’s “Cybernetic Poet” program can generate original stanzas of poetry. Where a generative technology’s behavior is emergent—complex and unanticipated by its creator—two interesting questions arise: (1) Should machine-generated art be entitled to copyright protection, and (2) if so, who gets the copyright?

In the United States, copyright protection “subsists . . . in original works of authorship” (Copyright Act of 1976, 17 U.S. Code § 102). An “original work” is one “independently created by the author (as opposed to copied from other works)” and that “possesses at least some minimal degree of creativity” (Feist Publications, Inc. v. Rural Telephone Service Co., 1991). This is “not a stringent standard” (ibid). For example, even mezzotint (a mechanical means of creating intermediate levels of grey by selectively roughening printing plates) copies of others’ famous paintings are “original works” entitled to copyright protection (Alfred Bell & Co. v. Catalda Fine Arts, 1951). “A copyists’ bad eyesight or defective musculature, or a shock caused by a clap of thunder, may yield sufficiently distinguishable variations,” which the copyist may adopt as his own and copyright (ibid).

Given this low standard, there is a strong argument that existing technologies like AARON and the Cybernetic Poet create “original works of authorship” entitled to copyright protection (Bridy, 2012). But who, if anyone, gets the copyright remains open for debate (for discussion, see Bridy, 2012; Clifford, 1997; Samuelson, 1985). The most obvious answer seems to be the human inventor (or programmer) of the machine (or program) that generates the art. The inventor is the human with the closest nexus to the art (Davis, 1991) and can arguably be treated like the generative machine’s employer under the work-for-hire doctrine (Bridy, 2012). But the fact remains that the inventor is not the author-in-fact of the art: The generative machine is the entity making the compositional choices. This observation has led to the interesting suggestion that autonomously-generated works should simply become part of the public domain, at least until technology advances to the point that granted copyright protection to a machine might make sense (Clifford, 1997). Of course, it is worth keeping in mind that we do not necessarily need a universal answer: Disputes could be resolved by courts on a case-by-case basis, perhaps based on whether the

---

8 Unlike patent protection, copyright protection does not require novelty or non-obviousness (for discussion, see Swinson, 1991; Alfred Bell & Co. v. Catalda Fine Arts, 1951).
“traditional elements of authorship in the work” at issue were “conceived and executed” by the machine or by the machine’s inventor (Register of Copyrights, 1966).

This issue remains open. When it is ultimately decided—by Congressional revision of the Copyright Act or by courts on a case-to-case basis—the resolution will likely be influenced by how the relevant decision-makers conceptualize generative technology. Similar to the Fourth Amendment context, we would expect that the more generative technologies are viewed as tools, analogous to paintbrushes, the more sense it makes to grant copyright protection to their inventors. Conversely, the more generative technologies are viewed as agents, the more appealing the public domain argument becomes (see Clifford, 1997). Thus, inventors who are concerned about copyright protection in the artwork their machines generate may want to be strategic in matters like name and design choices. Perhaps a name like “PaintBot” would be strategically preferable to a more anthropomorphic name such as “AARON.” If the generative machine is embodied, it may be wise to shy away from giving it a human form. Put in terms of the transition model, inventors may want to design generative machines such that cues relating to first-line processes (e.g., perceptual detection and classification) disfavor attributions of agency. Further, the transition model posits that anthropomorphism need not be all or nothing—people can (and do) limit attributions to particular domains. Thus, people may be surprisingly willing to consider a generative technology an agent for purposes of creating artwork, even if they do not consider it an agent for purposes of having legal rights in that artwork.

3.3. Statutory and regulatory interpretation

The U.S. Code, which includes all federal statutes, was 56,009 single-spaced pages long as of 2008 (Roberts & Stratton, 2008). The Code of Federal Regulations, which includes all of the rules and regulations promulgated by federal agencies, was 178,277 pages long as of the end of 2015 (George Washington Columbian College of Arts & Sciences Regulatory Studies Center, 2016). By the time you add in state and local codes and regulations, it seems safe to say there are more pages of law in the United States than anyone could ever read. Given this volume of law, there are, of course, more than a few ambiguities. And, even where provisions in these laws were not ambiguous when they were adopted, they can be rendered ambiguous by technological change.

For example, the vast majority of laws in the United States were not written with the widespread availability and use of robots in mind. Thus, as robots increasingly perform activities traditionally performed exclusively by humans, administrative agencies and courts will have to decide whether and how statutes and regulations governing those activities apply to robots. In Robots in American Law, Ryan Calo discusses several cases where this type of problem has already arisen (Calo, 2016). For instance, Comptroller of the Treasury v. Family Entertainment Centers (Md. 1987) addressed whether the animatronic bands that grace the stages of Chuck E. Cheese’s restaurants deliver “performances,” and thus subject the restaurants to an additional tax on food “where there is furnished a performance.” The court decided that the bands did not perform within the meaning of the statute, reasoning that a “pre-programmed robot can perform a menial task but, because a pre-programmed robot has no ‘skill’ and therefore leaves no room for spontaneous human flaw in an exhibition, it cannot ‘perform’ a piece of music” (ibid). While this rationale makes sense in the context of the Chuck E. Cheese band, it is an interesting question whether some more advanced modern robots, with their knack for adaptability and spontaneity, are capable of “performing.” In other examples, customs courts have addressed whether robots represent something “animate” for purposes of customs tariffs, and the National Highway Transportation Safety Administration recently indicated that in its view the word “driver” could include software in certain circumstances (Calo, 2016).

Questions of interpretation like these are perhaps especially prone to influence by the interpreters’ anthropomorphic attributions. To say a robot can deliver a musical performance, for example, is to acknowledge that technology has an ability that has previously been exclusively the province of humans—in that sense, the use of the label “performance” is itself a sort of anthropomorphic attribution. Naturally, then, we would expect that a judge who broadly anthropomorphizes a robot to be more likely to interpret a statute to include the robot’s actions. Similarly, we would expect that the actions of a broadly anthropomorphic robot—with a human form, a human name, and humanlike movements—to be more likely to fall within the scope of such statutes. In cases of interpretation like these, where anthropomorphic attribution is in a very real sense the heart of the issue, the transition model of agency suggests that the advocate for expanding the statute to apply to the intelligent technology (e.g., the party arguing that the robot gives “performances,” is “animate,” or is a “driver”) may want to do his or her best to induce cognitive dissonance in the relevant decision-maker.
On a broader level, the empirical work on attributions of agency highlights the importance of who is doing the interpreting. Put in the terms of the transition model, the availability of deep and/or broad concepts may induce transitions between first-line processing and second-line conceptualization—that is, how readily you can bring to mind information about the type of agent that you are reasoning about will influence your attributions. (Similarly, in the parlance of the promiscuous agency account, elicited agent knowledge tends to decrease attributions.) Thus, if the interpretative question of whether a robot can “perform” were being answered today, we would expect a regulator who is very familiar with robotics to anthropomorphize less (and thus be more inclined to answer the question “no”) than a generalist judge or a regulator unfamiliar with robotics. Depending on one’s policy views, this could provide support for the argument that a federal agency dedicated to issues involving intelligent technology and robotics is needed (Calo, 2014).

3.4. Negligence litigation

While data describing litigation patterns is hard to come by, the data we have indicates that negligence cases constitute the majority of tort litigation in the United States (see, e.g., Eaton & Talarico, 1996; Smith, DeFrances, Langan, & Goerdt, 1995; see also Goldberg, Sebok, & Zipursky, 2004 (“Because of its breadth, negligence is today the tort most commonly relied upon by accident victims to obtain redress from an alleged wrongdoer.”) (p.47)). In a negligence case, a plaintiff alleges that she has been injured in some way by the careless conduct of the defendant. To prevail on this claim (and thus recover compensation from the defendant), the plaintiff must demonstrate that four things are more likely true than not: (1) the plaintiff suffered an injury, (2) the defendant owed a relevant duty of care to the plaintiff, (3) the defendant breached that duty of care, (4) the defendant’s breach was both an actual and proximate cause of the plaintiff’s injury (e.g., Goldberg, Sebok, & Zipursky, 2004). Conventionally, the second element (duty) is considered a question of law for the court to decide while the other three elements are considered issues for the trier of fact (though, of course, a court can step in and rule on a dispositive motion if there are no factual allegations or disputes necessitating the involvement of a trier of fact). If the plaintiff fails to sufficiently prove any one of the four elements, then the plaintiff’s claim fails and she is entitled to no relief.

Over the years, the definitions, presumptions, and boundaries related to each element of a negligence claim have been refined through repeated applications, formulations, and re-formulations in cases. Today, if a plaintiff pedestrian walking along the sidewalk is (inadvertently) struck and injured by a distracted driver’s car, a body of precedent gives us confidence that the defendant will likely be liable to the plaintiff. Similarly, if a plaintiff pedestrian walking along the sidewalk is struck and injured by a rogue barrel falling out of the window of the defendant’s business, precedent gives us assurances that the defendant business will have to pay for the plaintiff’s injuries (see Byrne v. Boadle, 1863; Kambat v. St. Francis Hosp., 1997; Restatement (Second) of Torts § 328D, 1965; see also Pylman, 2009.) But, by and large, these elements have not yet been applied in (and have not been refined to accommodate) cases involving sophisticated, emergent machines—where the pedestrian is struck by a self-driving car (see Jones, Schall, & Shen, 2014) or where an autonomously-functioning robot (rather than a barrel) falls out of the window (see Calo, 2015). This section focuses on the impending prospect of cases like these.

Because plaintiffs cannot sue machines, the puzzle for law is figuring out who (if anyone) must pay when a machine’s emergent behaviors cause injury. Some scholars contend that contemporary negligence law can handle such a situation with few, if any, tweaks (Hubbard, 2014; Vladeck, 2014). Others posit that emergent behavior may catalyze the re-evaluation of at least some foundational elements (Calo, 2016). In our view, the extent to which emergent machines will prompt change to negligence law will depend on the extent to which decision-makers anthropomorphize the relevant machines. Our reasoning: If emergent machines prompt change, it will likely be to address cases where “a victim who suffers a non-natural harm [has] no perpetrator to whom the law can attribute this harm” (Calo, 2015). We anticipate that the prevalence of such cases will be positively correlated with decision-makers’ tendencies to anthropomorphize the machines at issue.

Specifically, attributions of agency could factor into the analysis of three elements of negligence: duty, breach of duty, and proximate causation. The duty element is a matter of law for the judge to assess (e.g.,

---

9 Interestingly, while we cannot sue machines, Peter Kahn and colleagues found that 65% of participants ascribed moral accountability to an android that deprived them of a $20 (Kahn et al., 2012).
generally speaking, a person owes others a duty to use reasonable care when engaging in conduct that could foreseeably harm them (ibid). If courts make anthropomorphic attributions to emergent machines, those attributions may tend to exculpate the owner or manufacturer, because the technology’s problematic behavior may be deemed unforeseeable (see Zipursky, 2009, n.47 (discussing relevance of foreseeability to duty analysis); see also Calo, 2015). With respect to breach of duty, one could imagine that—holding duty constant—a factfinder who broadly anthropomorphizes the relevant technology would be less likely to find breach that one who does not simply because, in his or her mental representation, there is one more “agent” in the scenario. Proximate causation provides another path for influence. Like duty, proximate causation focuses on foreseeability: Was the specific injury that befell the plaintiff a reasonably foreseeable consequence of the defendant’s actions? We posit that the more a finder of fact tends to anthropomorphize an emergent machine, the greater the likelihood that the injury it caused will be deemed unforeseeable to the defendant owner or manufacturer, thereby exculpating them.11

Because anthropomorphic attributions would tend to exculpate defendants, the promiscuous agency account implies that emergent machine negligence cases may cause a substantial disruption—that is, there may soon be a significant number of negligence cases in which plaintiffs are left with no remedy. The transition model of agency, however, forecasts less disruption. Due to the very nature of legal proceedings, many of the first-line processes that might invite anthropomorphic attributions are unlikely to be triggered. For example, if evidence about the relevant technology is introduced to the jury and/or the judge by a combination of descriptive testimony and photographs, and not by video or demonstration, then the sorts of cues that might trigger initial attributions of agency via perceptual detection and classification and basic agency/theory of mind and reference (e.g., the robot’s gaze patterns) will be unavailable. Similarly, because they are unlikely to interact directly with the emergent machine at issue, decision-makers are unlikely to apply social-cognitive defaults to it. If first-line processes do not favor anthropomorphic attributions, then, per the transition model, we would expect that such attributions would be difficult to trigger, occurring only as a result of deeper deliberation and reconceptualization on the part of the decision-maker. And, indeed, this expectation finds support in the empirical literature. Recall Levin and colleagues’ finding that, when making behavioral predictions, participants tended to equate robots with computers unless they saw the robot preferentially look at ten pairs of objects (a task that would trigger multiple first-line processes) (Levin, Killingworth, et al., 2013). When participants are presented with descriptions of a robots’ behavior (instead of experiencing it), they are often reluctant to anthropomorphize (e.g., Jaeger & Levin, in prep; Levin, Killingworth, et al., 2013; Levin, Killingworth, & Saylor, 2008).

The transition model of agency also has implications at a strategic level. Imagine you are a lawyer tasked with defending the owner or manufacturer of a recreational robot called “RoboBuddy X3” in a negligence case. You will want to understand whether it is advantageous to your client if the jurors think of the robot in anthropomorphic terms, and if it is (as we suggest above), you will want to try to maximize the likelihood that they do so. Perhaps you strategically refer to the robot as “Robbie” for short during proceedings. You may want to push—hard—to get demonstrative evidence about Robbie admitted. If the jurors do not see Robbie in action, you may want to get as much information about the robot and its functionality before the jury as you can (in an effort to facilitate deep consideration and reconceptualization). You could try to use testimony to link Robbie’s functions to concepts that are

10 Conceptually, this could be framed in terms of the “scope-of-the-risk” rule—the injury was outside the foreseeable scope of risk created by the owner’s or manufacturer’s actions (Medcalf v. Washington Heights Condominium Assn., Inc., 2000)—or, alternatively, it could be framed as the emergent machine functioning like a “superseding cause,” though that concept is falling out of favor (see Green, 2001).

11 These elements are treated as objective standards. For instance, with respect to duty, the issue would be whether a reasonable person would have foreseen the risks of the machine’s emergent conduct, not whether the machine’s manufacturer or owner actually foresaw them (or whether the judge would have foreseen them). Thus, it could be argued that the decision-maker’s personal attributions of agency should not affect their liability determinations. Nevertheless, we posit that they will. First, it strikes us as unlikely that a decision-maker considering an emergent machine case will stop to consider the nature of their attributions to the machine, let alone question whether they are the same attributions an objectively reasonable person would make. Further, even if people tried to do so, they would almost certainly still use their own experience and inferences as a guide in assessing what an abstract reasonable person would foresee (see, e.g., Epley, Waytz, & Cacioppo, 2007).
familiar and accessible to jurors. You could try to induce cognitive dissonance. The transition model suggests a variety of tactics that may be effective, none of which would be obvious without an empirical understanding of attributions of agency.

* * * * *

This section was intended to briefly illustrate the potential implications of anthropomorphic attributions—and, more specifically, of the transition model of agency—for a varied sample of robotics law issues. While the details of those implications vary from context to context, some common themes are worth noting. First, companies or individuals in the process of creating intelligent technologies may want to consider, among other things, the type of attributions their name and design choices invite (and the potential legal implications of those attributions). Choices relating to first-line defaults (e.g., designs that invite certain perceptual classifications) may have less-than-obvious consequences: A dancing robot with a face and two legs may be found to give “performances” for tax purposes, while the same robot with no face and four legs may not. Second, the transition model suggests that who decides questions related to intelligent machines—judges, legislatures, regulators, or perhaps a new specialized agency (Calo, 2014)—will likely matter for reasons beyond simple policy preferences. Empirical evidence suggests that greater expertise with the technology at issue will lead to fewer attributions of agency, which may, in turn, affect substantive policy. Third, when litigation arises with respect to intelligent technologies, the transition model suggests some strategies that may be useful for advocates. For instance, inducing cognitive dissonance in order to enhance attributions of agency could be effective in some circumstances. Finally, an overarching point embedded in the transition model (and reflected in the research) is that anthropomorphic attributions can be property- or function-specific (e.g., people who say that a technology that can consider alternatives will not necessarily say that it has consciousness). Thus, decision-makers’ conclusions about one dimension of agency may not entail similar conclusions about other dimensions, even if the dimensions are seemingly related.

4. Conclusion

As the American legal system prepares for an influx of issues relating to intelligent technologies, it stands to benefit from an empirical perspective. Our law and policy regarding such technologies will inevitably be shaped by decision-makers’ understandings of how they operate. Those understandings, in turn, are shaped in predictable ways by specific cognitive processes. Psychologists have studied these processes in considerable detail as they relate to interpersonal interactions, and an expanding body of research examines these processes in the context of human-robot interaction. This growing literature provides a sturdy foundation for addressing legally-relevant questions, such as those raised in our introductory example of the pill-dispensing robot. This literature can and should play a part in informing policy decisions concerning intelligent technologies. Ultimately, the legal system could use the firm grounding provided by a well-developed empirical account of anthropomorphic attribution—like the transition model of agency—to help weather the storm of technology-related issues that is gathering on the horizon.

Acknowledgments

This material is based on the work supported by the National Science Foundation under Grant No. 0826701 to DTL. Thanks to Ted Jaeger for reading and commenting on previous versions of this manuscript.

References

Alfred Bell & Co. v. Catalda Fine Arts, 191 F.2d 99 (2nd Cir. 1951).


Comptroller of the Treasury v. Family Entertainment Centers, 519 A.2d 1337 (Md. 1987).


Restatement (Second) of Torts § 328D (Am. Law. Inst. 1965).


Corresponding Author’s Contact Information: C. B. Jaeger, Department of Psychology and Human Development, Vanderbilt University, Peabody College # 552, 230 Appleton Place, Nashville, TN 37203-5701. Email: christopher.b.jaeger.1@vanderbilt.edu