The Dissertation Committee for Gabriel Lopez-Mobilia Certifies that this is the approved version of the following dissertation:

Children's Psychological and Moral Attributions to a Humanoid Robot

Committee:

Jacqueline D. Woolley, Supervisor

Catherine H. Echols

Arthur B. Markman

Lauretta Reeves

Peter H. Stone

Children's Psychological and Moral Attributions to a Humanoid Robot

by

Gabriel Lopez-Mobilia, B.S.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2015

Children's Psychological and Moral Attributions to a Humanoid Robot

Gabriel Lopez-Mobilia, Ph.D.

The University of Texas at Austin, 2015

Supervisor: Jacqueline D. Woolley

In the near future, sophisticated social robots will become increasingly interwoven into our lives. Researchers have recently begun to examine people's anthropomorphic conceptions of such robots, and a few have stressed the unique consequences that these technological agents may have for the psychological development of children developing around them. In the current set of studies, children were introduced to a humanoid robot, "Robbie the Robot." Across the two studies, participants witnessed Robbie perform a harmful action, destroying a block tower that a child had purportedly built and was saving for later. Of primary interest in these two studies was whether children would hold Robbie the Robot morally accountable for the destructive act. It was predicted that judgments of moral accountability would depend on several different factors: whether the robot appeared to initiate its own actions, the age of the participant, and whether children attributed psychological properties, specifically intentional agency, to the robot. In Study 1, children were assigned to one of two experimental conditions: a *controlled* condition in which a confederate appeared to control the robot's actions with a device that was tethered to the robot, and an

autonomous condition in which the robot appeared to move of its own accord. Results revealed that children were significantly more likely to attribute psychological properties to the robot in the autonomous condition compared to the controlled condition. Compared to 7-year-olds, 5-year-olds were more likely to attribute psychological properties to the robot overall. In addition, results indicated that increasing cues to the robot's autonomy indirectly affected moral accountability judgments through an increase in children's attributions of intentions. Study 2 tested the hypothesis that children's attributions of psychological agency, but not psychological experience, would increase after watching the robot commit a moral act. Overall, Study 2 results did not support this prediction, but key results from the first study were replicated and elucidated by the inclusion of a wider array of psychological properties as well as a measure of children's judgments of the robot's cuteness. Implications are discussed for human interaction with social robots and other rapidly evolving technologies, such as autonomous vehicles.

Table of Contents

List of Tables

List of Figures

Introduction

For at least a century, writers of science fiction have dreamed of humanlike machines invading the lives of people on Earth. Now, within our lifetimes, we are beginning to experience the emergence of increasingly sophisticated social robots capable of interacting with their physical and social environments autonomously and creating convincing illusions of humanlike personalities. How will we receive such robots as they enter into our lives? Will we treat them with similar compassion that we grant other humans or our pets? Will we be willing to place our trust in them? Will we hold them morally accountable for their actions? How will young children come to conceptualize such robots as they develop around them?

Since its inception in the 1940's and '50's (Walter, 1950), the field of social robotics has flourished, spurred by research in artificial intelligence emphasizing the creation of autonomous humanlike robots that can effectively interact with and communicate with humans. In the past few decades, researchers have made great strides in developing robots capable of a range of humanlike features for enhancing social interaction with humans, including the ability to identify and express emotions through facial gestures (e.g., Breazeal & Scassellati, 1999), the ability to exhibit shared attention with humans (Scassellati, 2002), and the ability to imitate actions (e.g., Mataric, 2000). Such robots are also being developed for use as companions and caretakers for the elderly (Wada & Shibata, 2007; Roy, Baltus, Fox, Gemperle, Goetz, Hirsch, Margaritis, Montemerlo, Pineau, Schulte, & Thrun, 2000) and as tools for diagnosing and potentially

treating children with autism (Scassellati, 2007; Robins, Dautenhahn, Te Boekhorst, & Billard, 2005). Although many of these state-of-the-art robots are still in development, various humanoid and animal robots have already become commercially available, especially for entertainment (Kahn, Gary, & Shen, 2013). With falling costs, more advanced social robots that can be used for companionship and caregiving of young children and elderly individuals are likely to become common in the near future (Sharkey & Sharkey, 2011).

With the recognition that these sophisticated personified technologies will become increasingly interwoven into people's lives, researchers have begun to address important questions regarding how people conceptualize humanoid robots, including under what conditions people ascribe moral status to them and in what circumstances people find them unnerving (as with the "uncanny valley"; Gray & Wegner, 2012), and a few have even stressed the unique consequences that these robots may have for the psychological development of children growing up around them (Kahn et al., 2013; Severson & Carlson, 2010; Sharkey & Sharkey, 2011). While robots have the potential to assist people in daily activities and provide companionship, their unique status as both a social entity and a mechanical artifact poses a unique challenge with regard to conceptualizing their psychological and moral status (Kahn et al., 2013). Investigating children's interactions with social robots can shed light on these important issues while providing a unique context for studying the cognitive-developmental underpinnings of anthropomorphism.

The current study will examine just a few of the conditions that predict whether a child will anthropomorphize a robot, granting it humanlike mental states and moral status. I will begin by introducing the concept of anthropomorphism, reviewing some of the factors that are known to lead to the attribution of humanlike psychological traits to nonhuman entities. This will lead into a discussion about the moral consequences of attributing a mind to an entity. Then, I will review previous work on children's and adults' psychological and moral attributions to robots specifically. Finally, I will describe findings from two experimental studies that were designed to investigate relations between perceived agency in a humanoid robot and young children's attributions of psychological and moral status to it, and I will discuss implications and future directions.

Literature Review

ANTHROPOMORPHISM

The term *anthropomorphism* (from the Greek words *anthropos*, meaning "human", and *morphe*, meaning "form") refers to any attribution of human characteristics to nonhuman entities. In the current paper, the term *anthropomorphism* will refer specifically to the attribution of humanlike psychological (or mental) states to real or imagined nonhuman entities (Epley, Waytz, & Cacioppo, 2007). These mental properties include desires and motivations, emotions, and the ability to think or act intentionally.

Although robots are a modern-day phenomenon, anthropomorphism is ancient. The tendency to attribute human properties to nonhuman entities may date back to the Upper Paleolithic (Mithen, 1996), and it appears to be a cultural universal (Hume, 1757/1957; Guthrie, 1993). Indeed, one of the oldest sculptures ever created, thought to be about 32,000 years old, depicts a form with the body of a human and the head of a lion (Mithen, 1996). In the modern world, people tend to anthropomorphize a wide rage of entities that they regularly think about and interact with including nonhuman animals, technological devices, and nature. Within psychology there has been a recent surge of interest in understanding the social-cognitive underpinnings of anthropomorphism. In particular, research is beginning to uncover factors that lead people to anthropomorphize and the consequences of such anthropomorphism (Epley et al., 2007).

Anthropomorphism as Social Cognition

Various researchers (e.g., Epley et al., 2007; Bloom, 2004; Kwan, Gosling, & John, 2008) have proposed that anthropomorphism is fundamentally related to social cognition. Under this account, when anthropomorphism occurs, cognitive mechanisms that are typically dedicated to reasoning about humans are used for thinking about nonhuman entities. In fact, some existing evidence suggests that perceiving both human and nonhuman intentional actions or mental states involves the activation of similar brain regions (Waytz, Morewedge, Epley, Monteleone, Gao, & Cacioppo, 2010; Gazzola, Rizzolatti, Wicker, and Keysers, 2007). For example, in one study by Gazzola et al. (2007), participants were scanned using fMRI while viewing simple actions performed by either a human hand or an industrial robotic hand (e.g., placing a lid on a jar). The researchers were interested in examining the activation of brain areas typically associated with the activity of mirror neurons, that is, neurons that fire both during the execution of an intentional action and during the observation of that same action performed by another individual (Rizzolatti & Craighero, 2004). Results showed that the actions, especially actions that were goal-directed, resulted in the activation of areas typically associated with the mirror neuron system regardless of whether they were performed by the human hand or the robotic hand. Thus, anthropomorphism may be partly underpinned by a mirror neuron system that flexibly responds to a relatively broad range of stimuli. This flexibility may result in the processing of nonhuman entities in social terms.

In addition, there is evidence that impairments in social cognition (e.g., in autistic individuals) are linked to impairments in the attribution of mental states to nonhuman

entities (Gray, Jenkins, & Heberlein, 2011; Castelli, Frith, Happé, & Frith, 2002). For example, Castelli et al. (2002) compared the brain activation of autistic and non-autistic adults while they viewed sequences of animations portraying geometric shapes moving about in ways that suggested intentional behaviors. Results revealed that the autistic individuals were less successful in accurately describing the mental states implied by the movements of the shapes, and PET scans showed that the autistic group had significantly less activation in areas typically associated with the understanding of mental states including the medial prefrontal cortex (MPFC).

As a special case of social cognition, anthropomorphism may also be conceptualized as a specific application of theory of mind (sometimes called mentalizing) to think about nonhuman targets. Theory of mind is the ability to reason about other minds by making inferences about individuals' unseen beliefs and desires that are different from one's own point of view. In the developmental psychology literature there are two prominent competing ideas about the development of this kind of social cognition: Theory Theory and Simulation Theory. According to Theory Theory, children develop coherent bodies of knowledge for understanding distinct domains in the world, including the physical, biological, and social domains. On this account, theory of mind emerges as a folk psychological framework for representing the mental states of others and ourselves (Perner, 1991). In contrast, Simulation Theory proposes that understanding the mental states of others is achieved by using our own minds and bodies to model and predict others' mental states based on how we would act under similar circumstances (Gallese & Goldman, 1998; Harris, 1992). At the neural level, simulation may occur

through the activation of the mirror neuron system (Gallese & Goldman, 1998). By directly mapping the observed behaviors of other individuals onto the same system that coordinates the execution of similar behaviors, the mirror neuron system may in effect create a platform for the simulation of other minds.

Development of Anthropomorphism

If anthropomorphism can be traced to social cognition, an examination of its development should begin with an understanding of the development of social cognition, from attributing intentional agency to having a theory of mind. Detecting other people and reasoning about their behavior is a vital task for surviving in the ultra-social world of human beings from the earliest stages of life. Soon after birth, an infant's gaze is drawn toward visual configurations resembling the configuration of a face (i.e., top heavy elements bounded by a margin; Johnson & Morton, 1991). Within a few months of life, infants recognize and discriminate human biological motion portrayed with mere moving light points (Bertenthal, 1993).

By about 6 months, infants appear to represent the actions of certain agents as goal-directed (Carey, 2009). Between 6 and 12 months of age, self-propelled motions of geometric figures and their contingent movements with respect to other shapes in a scene are enough to elicit rich interpretations and predictions about the actions of goal-directed agents (e.g., Gergely, Nadasdy, Cisbra, & Biro, 1995) and about their dispositions (e.g., Hamlin, Wynn, and Bloom, 2007). For example, in one study, Hamlin et al. (2007) showed 6- and 10-month-old infants a scene with a geometric shape (e.g., a circle) with

eyes "attempting" to climb up a hill. Then, infants were shown instances where another shape (e.g., a triangle with eyes) "helped" the climber by pushing it up and instances where yet another shape (e.g., a square with eyes) "hindered" the climber by pushing it down the hill. After viewing these events, infants were presented with physical objects shaped like the helper and the hinderer, and reaching behaviors were interpreted as the infant's preference for one of the two. Results revealed that infants reliably preferred to reach for the shape that had been previously portrayed as a "helper." This study and more recent studies by Hamlin and colleagues (e.g., Hamlin & Wynn, 2011) provide compelling evidence that infants as young as 5 months old are not only making clear distinctions between animate and inanimate objects in the world but that they are also beginning to attend to potential social cues to make rich dispositional (and possibly moral) inferences about animate entities.

By the age of 18 months, toddlers have become keenly attuned to the intentions behind people's actions, and they begin to use these cues to engage in prosocial behaviors (Warneken & Tomasello, 2006). For example, in one experiment, 18-month-olds observed as an adult "attempted" but failed to complete a goal such as reaching for a dropped marker or attempting to place a book on top of a stack. After observing the adult having trouble, infants reliably inferred that the adult needed help and spontaneously helped to complete the goal despite never having seen the action through its completion.

By the age of 3 years, children exhibit a full-blown schema of intentional causality, a theory of mind, invoking beliefs and desires to explain the behaviors of different agents. Around the age of 4 years, children begin to reliably pass false-belief

tasks, explicitly making accurate predictions about others' mental states that diverge from their own beliefs (Wimmer & Perner, 1983). This ability to accurately reason about other minds continues to develop throughout the preschool years (Bartsch & Wellman, 1995). In sum, early on in development, humans are keenly attuned to cues to animacy in agents and view agents as intentional and goal-directed, distinguishing them sharply from inanimate objects. This basic distinction between animate and inanimate appears to form the basis for the attribution of a mind that is capable of psychological experiences and as such would appear to form the foundation for anthropomorphism.

In contrast to the large body of work on social cognition and theory of mind about other humans, until recently there has been a limited amount of research directly and systematically examining anthropomorphism in children. Nonetheless, since Piaget (1929) it has commonly been claimed that young children are highly animistic, indiscriminately attributing life and psychological properties to nonhuman entities like clouds and trees, and that claim has often been extended to imply that children are rampant anthropomorphizers (e.g., Epley et al., 2007). However, Piaget's original claim of animistic children has more recently been challenged by research with careful systematic methods showing that even children as young as 3 years old hold a firm understanding of the animate-inanimate distinction (e.g., Carey, 1985; Jipson & Gelman, 2007), implying that the assumption that children are indiscriminate anthropomorphizers is probably unwarranted.

Nonetheless, studies suggest that young children may hold some anthropomorphic ideas about nonhuman entities such as plants, sometimes claiming that they can have

psychological experiences such as pain. In one study by Inagaki and Hatano (1987), 5- to 6-year-old children were asked to reason about nonhuman entities (e.g., a tulip, a rabbit) and to predict their behavior in different scenarios. Results showed that children were likely to personify the entities through analogy with humans in situations in which this allowed for a reasonable prediction of the entity's behavior. For example, when asked about what would happen to a tulip if was not given water for days, children responded that it would feel thirsty, sometimes referring to a similarity to humans. The researchers also found that although children were quite good at identifying observable (physical) attributes of the entities, they often attributed unobservable (psychological) attributes, with a substantial number of children even claiming that tulips feel happy, feel pretty, feel pain, and feel cold (Inagaki & Hatano, 1987). If children are in fact competent at categorizing animate and inanimate entities, why do they sometimes make such striking anthropomorphic claims?

A study by Jipson and Gelman (2007) may provide some clarification to this seemingly inconsistent account of young children's proficiency in understanding nonhuman entities. The researchers asked 3-, 4- and 5-year-olds and adults about the biological and psychological properties of different entities (e.g., a robot, a starfish) that varied on three dimensions: whether the entity was alive, whether it had a face, and whether it exhibited autonomous behavior. Results indicated that even children as young as 3 years old made a firm distinction between living and nonliving objects in terms of possessing biological properties. However, the distinction between living and non-living did not govern young children’s attributions of psychological properties, especially with

items that had a face (e.g., a robot), suggesting that "for children, items can be nonbiological, but psychological" (Jipson & Gelman, 2007, p. 1686). Thus, at very young age, children are able to accurately distinguish living from non-living entities, but their attributions of psychological traits are nonetheless influenced by salient cues that are commonly associated with animate entities.

Anthropomorphism as Magical Thinking

Because there is consensus that nonhuman entities are not typically actually capable of exhibiting humanlike mental states, anthropomorphism can be considered a form of magical thinking or a fantasy-reality confusion (Hutson, 2012). Indeed, a recent study by Willard and Norenzayan (2013) shows that anthropomorphism in adults predicts other forms of supernatural thinking, such as the endorsement of paranormal beliefs. Like other forms of magical thinking (Woolley, 1997), the tendency to anthropomorphize may not be fundamentally different across development, and in certain contexts adults may be just as prone to anthropomorphism as children.

Interestingly, in Jipson and Gelman's (2007) study, even adults' psychological attributions appeared to be slightly increased by the presence of a face in an entity. This result is consistent with recent work showing that adults often anthropomorphize nonhuman entities (Waytz, Cacioppo, & Epley, 2010), and it is in accord with the idea that adults are not fundamentally different thinkers than children in their engagement with various forms of magical thinking (Woolley, 1997). Instead, children and adults alike may be prone to making relatively automatic anthropomorphic attributions given

the appropriate triggers, but these attributions may become suppressed at an explicit level as children develop the executive functions necessary to override them.

Such a pattern has been proposed to exist in many domains of magical thinking, with adults merely overriding natural tendencies to think magically and the tendencies resurfacing in certain contexts (Hood, 2008; Subbotsky, 1993). For example, in one study, Keleman, Rottman, and Seston (2013) showed that, when cognitive resources were limited through a speeded response task, adults were likely to default to teleological explanations typical of children, agreeing with the veracity of purpose-based explanations for natural phenomena (e.g., "The sun radiates heat because warmth nurtures life"). Other work has shown that elderly individuals with Alzheimer's tend to exhibit animistic thinking, endowing inanimate objects with attributes of living agents (Zaitchik & Solomon, 2008). These findings are consistent with the idea that the magical thinking exhibited by young children does not altogether disappear in adults, rather, it may become suppressed at an explicit level and resurface with the deterioration of the cognitive resources necessary to inhibit them. Anthropomorphism may be a phenomenon that exhibits a similar pattern to this, but more research is needed to examine this possibility.

In sum, anthropomorphism appears to share some fundamental structure with other forms of magical thinking, and it appears that proneness to magical thinking may even be one factor that predicts whether an individual will anthropomorphize. However, a number of other factors have also been shown to contribute to anthropomorphism. As a

form of social cognition, anthropomorphism appears to be triggered primarily by the same cues that are responsible for person perception.

Predictors of Anthropomorphism

In a recent review article, Epley et al. (2007) put forth a three-factor theory for explaining and predicting anthropomorphism. Approaching anthropomorphism as a form of inductive inference, the researchers outline three primary determinants of when people will perceive humanlike minds: sociality motivation, effectance motivation, and elicited agent knowledge. The three-factor model for anthropomorphism is useful for predicting when people will anthropomorphize a range of entities across situational and dispositional contexts.

Sociality motivation has to do with the tendency for humans to seek out social agents and to avoid being alone. When sociality motivation is high, individuals are more likely to detect humanlike entities. For example, in one study by Epley, Waytz, Akilis, and Cacioppo (2008), adults who reported higher levels of loneliness were more likely to anthropomorphize their pets, ascribing them higher levels of psychological attributes relating to social connection (e.g., rating them as thoughtful or considerate). In another study, Epley, Akilis, Waytz, and Cacioppo (2008) found that participants who were primed to feel socially disconnected were more likely than comparison groups to anthropomorphize their pets and to report belief in supernatural agents. Thus, the desire for social connectedness appears to increase peoples' tendency to anthropomorphize.

Effectance motivation refers to people's desire to interact effectively with their environments (White, 1959). With respect to agents, effectance motivation is high when there is a desire to control or predict behavior. According to Epley et al. (2007), anthropomorphizing an unfamiliar entity may be one way to increase its perceived predictability by ascribing to it a familiar humanlike agency. Supporting this idea, Waytz, Morewedge, Epley, Monteleone, Gao, and Cacioppo (2010) demonstrate in a series of studies that when adults are primed to think of computers or gadgets as unpredictable in their behavior, they are more likely to anthropomorphize them, rating them as "having a mind of its own," and having "intentions, free will, consciousness" (p. 415).

Finally, elicited agent knowledge, refers to the accessibility and applicability of knowledge about humans when making inferences about nonhuman entities (Epley et al., 2007). As described earlier, anthropomorphism appears to draw upon a wide range of mechanisms responsible for social cognition. When the same cues that trigger inferences about humans (e.g., eyes, self-propelled motion, etc.) are exhibited by a nonhuman agent, a perceiver is likely to attribute humanlike mental states. It is worth noting that roboticists exploit this kind of elicited agent knowledge by designing robots with such cues that create convincing illusions of intelligence (Zawieska, Du, & Sprońska, 2012).

Relations Between Anthropomorphism and Moral Reasoning

Perhaps one of the most important consequences of anthropomorphism is that it may be related to the moral treatment of nonhuman entities. With respect to judgments about other humans, we know that children and adults take into account a person's

psychological experiences and intentions when making moral judgments about them (Killen & Smetana, 2013). For example, in one study by Zelazo, Helwig, and Lau (1996), children's moral evaluations were elicited for scenarios that involved normal causality (e.g., hitting causes pain) or an unusual causality (e.g., hitting causes pleasure). Results revealed that, regardless of causality, children as young as 3 years olds made moral judgments based on the internal experience of the victim; if the victim felt pain, then the action was condemned and deemed worthy of punishment. Thus, when a nonhuman entity is anthropomorphized and seen as capable internal experience such as pain, it may become subject to the same kind of social cognitive reasoning in the moral realm.

Indeed, Waytz, Cacioppo, and Epley (2010) provide some evidence that anthropomorphism is related to granting moral status to nonhuman targets. These researchers created an "individual differences in anthropomorphism questionnaire" (IDAQ) and demonstrated that adults have stable tendencies to anthropomorphize a range of nonhuman entities. The IDAQ includes items from three categories (natural entities, technological devices, and nonhuman animals), and participants rate the extent to which these items exhibit human-like psychological attributes (e.g., intentions, free will) on a scale from 0 (not at all) to 10 (very much; e.g., "To what extent does the average robot have consciousness?"). A high IDAQ score indicates a high tendency to anthropomorphize. In a series of studies, the researchers found that scores on the questionnaire predicted attitudes and behaviors toward nonhuman agents including "the degree of moral care and concern afforded to an agent, the amount of responsibility and trust placed on an agent, and the extent to which an agent serves as a source of social

influence on the self" (Waytz et al., 2010, p. 219). Thus, individual differences in the general tendency to anthropomorphize seem to predict moral judgments about nonhuman entities, including moral concern for them and judgments about their moral responsibility.

In addition to an examination of how moral judgments are affected by anthropomorphism in a general sense, it is also important to consider the specific *kinds* of minds that people are likely to see in the world around them (Dennett, 2008) and the specific moral attributions that may be tied to different kinds of minds. Previous work on mind perception with adults (Gray, Gray, & Wegner, 2007) suggests that people do more than simply attribute or fail to attribute a mind to an entity; rather, they tend to perceive distinct kinds of mental qualities (e.g., the ability to plan versus the capacity to feel) relatively independently of one another and at varying degrees along a continuum. Furthermore, different kinds of mental attributions appear to have different consequences for the treatment of an entity.

For example, in the study by Gray et al. (2007), adult participants compared the mental capacities of various entities including a tree, dog, robot, human adult, human baby, and God. The researchers calculated mean values for each mental capacity and a factor analysis revealed a solution that loaded onto two principle factors they called *Experience* and *Agency*. Experience accounted for 88% of the variance in the model and included hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy, and *Agency* accounted for 8% of the variance and included self-control, morality, memory, emotion recognition, planning, communication, and thought.

Importantly, while both Agency and Experience correlated with greater liking and concern for an entity, there were some important moral distinctions. Whether an entity deserved punishment for wrongdoing correlated more strongly with Agency than with Experience, and the desire to protect an entity from harm correlated more with Experience than with Agency.

Thus, the distinction between different kinds of minds appears to have moral implications, such that entities with greater agency tend to be perceived as more morally responsible and less likely to be victims in moral transgressions, while entities with greater experience tend to be seen as less morally accountable and more susceptible to becoming a victim (Gray & Wegner, 2009). Indeed, moral discussions about the protection of certain entities (e.g., fetuses or nonhuman animals) often hinge on whether the entity in question can accurately be attributed the capacity for psychological experience, especially the experience of suffering (Bentham & Browning, 1843; Dennett, 2008). For many entities in the world, even scientists and philosophers continue to disagree widely over what kinds of psychological states of consciousness can be accurately attributed. For example, with regard to the mental lives of animals, in the 17th century Rene Descartes argued that only humans possess consciousness, and academic treatment of the subject since has been characterized by debates about where to draw the line between what entities should and should not be considered conscious (Burghardt, 1985). The moral implications of where one draws this line between conscious and not conscious are so great that, in an attempt to publicly resolve some of the ambiguity regarding non-human animals, on July 7, 2012 a prominent group of neuroscientists

convened to sign "The Cambridge Declaration of Consciousness," stating that all mammals, all birds, and many other creatures should be considered capable of exhibiting states of consciousness (Low, Panksepp, Reiss, Edelman, Van Swinderen, Low, & Koch, 2012).

Although the artificial intelligence exhibited by current-day robots is generally considered to be a mere imitation of human consciousness, whether legitimate machine consciousness may one day be achieved is still an open question. Some philosophers such as David Chalmers (1996) have argued that a computer system capable of performing the right kinds of computations can theoretically be considered to exhibit states of consciousness. Regardless of whether experts ultimately come to a solid consensus about this philosophical issue, with increasing sophistication robots will certainly continue to blur the line and convince many of their sentience. In particular, children who develop around social robots and become psychologically attached to them will likely form legitimate beliefs that such robots have real mental lives (Severson & Carlson, 2010), making them subjects for moral consideration.

CHILDREN'S PSYCHOLOGICAL AND MORAL ATTRIBUTIONS TO ROBOTS

Previous studies on children's social relationships with animal robots and humanoid robots have found that children tend to attribute at least some psychological properties and moral status to them. For example, in one study Melson, Kahn, Beck, and Friedman (2009) compared children's interaction with the robotic dog, AIBO, and a live dog (an Australian Shepherd). Although children tended to attribute greater amounts of

psychological and biological properties to the live dog, children also engaged socially with the robotic dog and a majority affirmed that AIBO had mental states (e.g., it can feel embarrassed), sociality (e.g., AIBO could be their friend), and moral standing (e.g., it would not be ok to harm AIBO).

Although research has shown that children do, in fact, tend to attribute certain psychological properties and moral status to lifelike robots, there may be nuanced differences across development. For example, younger children generally tend to attribute more psychological properties to robots compared to older children (e.g., Kahn, Kanda, Ishguro, Freier, Severson, Gill, Ruckert, & Shen, 2012), but conceptions of robot intelligence also seem to change qualitatively between 5 and 7 years (Scaife & van Duuren, 1995). It appears that older children's attributions of intelligence to robots stem partly from an understanding that robots are controlled by computers and that the functions of a computer may be referred to as intelligent in the sense that they are similar to the functions of a brain. An interesting question is whether this kind of shift in the conception of a robot's psychological status is likely to be associated with a shift in moral consideration for the robot as well, given the relation between particular kinds of psychological attributions and moral status.

Some previous research has investigated how older children and adults conceptualize the moral status of robots. In study by Kahn et al. (2012), 9-, 12-, and 15-year-olds who engaged socially with a humanoid robot named Robovie, attributed some degree of moral standing to it, claiming that it had feelings and could be a friend, but denied that it should be granted certain rights such as liberty. In another study, performed

in the same lab, participants were likely to hold Robovie morally accountable for its actions after the robot "cheated" the participant out of winning a cash prize (Kahn, Kanda, Ishiguro, Gill, Ruckert, Shen, Gary, Reichert, Freier, & Severson, 2012). Participants' responses also indicated that they viewed the robot as being more morally accountable for its actions than a vending machine but less morally accountable than a human under similar circumstances. The results of these studies suggest that people may grant robots some kinds of moral status while denying them others. In addition, it is likely that the particular kinds of moral status granted to a robot may depend on the particular mental states attributed to it (Gray, Young, Waytz, 2012).

In the work on dimensions of mind perception by Gray et al. (2007), participants rated robots as having relatively high Agency and low Experience, the dimensions that they also found to predict specific moral conceptualizations. However, the authors of the study do not report whether individual differences in these attributions to robots predicted their moral ratings to the robot item specifically. Nonetheless, a logical conclusion of their findings would be that robots are more likely to be held morally accountable for their actions while being granted relatively lower degrees of moral standing. However, specific cues offered by specific robots may lead to relative increases or decreases in attributions of Agency and Experience which may lead to different evaluations of a particular robot's moral status.

One such cue that has been examined in at least a handful of experiments with young children and adults is whether a robot's actions appear to be self-initiated or remote-controlled (Gary, 2014; Gary & Chernyak 2013; Somanader, Saylor, & Levin,

2011). As discussed earlier, self-propelled motion is a powerful cue that leads to attributions of agency (Gergely et al., 1995). Indeed, Somander et al. (2011) found that children made more psychological attributions to a relatively simple humanoid robot when it appeared to move on its own compared to when the experimenter appeared to control its actions by pressing buttons. In addition, Gary and Chernyak (2013) found that 7-year-olds, but not 5-year-olds, were more likely to grant a robotic dog (AIBO) moral standing (i.e., concern for its wellbeing) when its actions seemed autonomous than when the experimenter appeared to initiate the robot's actions with a videogame controller.

Researchers have yet to examine the effects of autonomous motion cues on whether a robot would be held morally accountable for actions that could cause potential harm. Because the autonomous initiation of behaviors is directly related to planning and acting, cues that indicate self-initiated action may actually lead to a relatively greater increase in the attribution of psychological agency (e.g., intentions) compared to psychological experience (e.g., emotions). Thus, cues that suggest autonomy may also lead to a greater increase in the judgment of a robot as morally accountable relative to judgments that it should be regarded with moral concern.

In the current set of studies, the aim was to assess the effects of such autonomy cues on young children's psychological and moral attributions to a humanoid robot following a live interaction. Critically, children were asked to evaluate the robot's moral accountability for committing a harmful act. In addition, children's attributions of psychological and biological traits to the robot were elicited before and after the harmful

act, and a questionnaire was administered to parents in order to assess effects of children's prior exposure to robots.

Method of Study 1

PARTICIPANTS

Participants were 34 5-year-olds (mean age = 5;6, range = 5;1-6;0; 21 boys and 13 girls), and 33 7-year-olds (mean age = 7;5, range = 7;0-7;11; 15 boys and 18 girls). Race of participants was primarily Caucasian. Five additional children were tested but excluded due to procedural error, namely due to the robot malfunctioning. Participants were recruited from the database at the UT Children's Research Lab.

MATERIALS

The Robot

The humanoid robot used in the current studies was a *Nao*, designed by Aldebaran Robotics for friendly social interaction. The Nao robot is 58 centimeters (approximately 2 feet) tall and its physical appearance consists of rounded, cute, features (see Figure 1). It is capable of humanlike bipedal motion, and it can sit down and stand up on its own. The Nao includes a range of other humanlike features such as speech capabilities, sensors for vision and hearing, and hands for manipulating objects, but these features were not utilized in the current experiment. During the study, the robot was referred to as "Robbie the Robot."

Property Attributions Task

Stimuli consisted of five laminated cards, each with a color photograph of a different entity, including a baby, a tree, a car, a computer (Figure 2), and Robbie the

Robot (see Figure 1). Children responded to property attribution questions using a 4-point scale that consisted of a large thumbs-down ("no, not all"), a small thumbs-down ("no, not much") a small thumbs-up ("yes, a little"), and large thumbs-up ("yes, a lot"; see Figure 3).

Parent Questionnaire

Before the study began, parents filled out a questionnaire about their child's previous experience with robots. The first item asked "Has your child ever seen a robot in person before today?" They rest of the items were rated for frequency on a 5-point scale (1 = "Rarely / Never," 3 = "On occasion," and 5 = "Very often"): "Has your child played with robotic toys that resemble animals or people?"; "How often has your child watched movies/shows with robot characters, read books with robot characters, or played video games with robot characters?"; "Have you ever talked to your child about how robots work?"; and "How often does your child interact with dogs, cats or other animals?"

PROCEDURE

Two experimenters were involved in the study protocol. After parents and children gave their consent to participate in the study, the first experimenter (E1) escorted the child into a small room where s/he was introduced to the Robbie the Robot, sitting on the floor in the corner of the room, and to a second experimenter (E2) sitting in a chair nearby. Throughout the study, children observed the robot performing a series of actions, but critically, they were assigned to one of two conditions: a Controlled condition or an

Autonomous condition. In the Controlled condition, E2 appeared to control the robot with a controller that was tethered to the robot with a green cable, claiming to cause each of the robot's actions as they occurred (e.g., "Look! I'm making Robbie stand up"). In contrast, in the Autonomous condition, the robot appeared to be self-propelled with no visible external control, and E2 merely narrated the robot's actions (e.g., "Look! Robbie's standing up") immediately after the initiation of each action. In the first part of the study, children observed as Robbie the Robot stood up, stretched, walked forward towards the child, stopped and waved in the direction of the child, and finally sat down.

After observing these behaviors, children were escorted to an adjacent room where they were asked to sit at a table to answer some questions. Children were introduced to the 4-point scale that they would use to answer the questions, and they warmed up with three practice questions in which they were asked to rate whether they liked: 1) candy 2) broccoli, and 3) carrots. After the warm up, children were asked to rate how much they liked Robbie the Robot. Then, the property attribution trials began. The experimenter shuffled the entity cards (baby, tree, car, computer, and Robbie the Robot), displayed the backs of them and asked children to select the first one at random. After flipping over the card to display the image of the entity, children were told that the following questions would all be about that entity.

Children were asked to rate each entity with a series of property attribution questions. The order of questions was randomized before the before the start of the trials, and that same order was maintained for each entity. Psychological attribution questions included two agency items, one regarding *thoughts*: "Can [entity] think?"; and the other

about *intentions*: “Can [entity] do things on purpose?”; and two experience items, one pertaining to *emotions*: “Can [entity] feel things like happy or sad?”; and one about *sensations*: “If someone poked [entity], would [entity] feel it?” Biological questions were “Can [entity] grow?” and “Does [entity] eat?” Two questions were about whether the entity was worthy of moral concern: “Would it be OK to yell at [entity]?” and “If someone was mean to [entity], could [entity] get upset?” There was also one question about whether the entity was an artifact: “Did someone build [entity]?”

After property attributions were elicited for each of the five entities, children were escorted back into the room with Robbie the Robot. After children sat down, their attention was directed toward a tower on the floor made from colorful toy blocks, and they were told that a girl/boy (gender matched to participant) had built the tower earlier that day and that the tower was being saved so that the girl/boy could come back and show their friends later. Right after this back-story, the robot stood up and walked over to the block tower. After a brief pause, the robot proceeded to destroy the tower with a punch of its fist. In the Controlled condition, E2 continued to narrate control of the robot’s actions as they occurred (“Look, I’m making Robbie stand up” and “I’m making Robbie walk”) whereas in the Autonomous condition, immediately after the initiation of each action, E2 stated, “Look, Robbie is standing up” and “Robbie is walking.” After the robot knocked over the tower[1], children were escorted by E1 back to the interview room,

and doors were kept closed to insure that participants felt a sense of confidentially during the moral accountability interview that followed.[1]

Children were asked a series of questions about the severity of the event and about the involvement and moral accountability of Robbie the Robot and of E2. First, children were asked to explain what had just happened, and then they were asked, "Do you think what happened with those blocks was OK or was that not OK?" If they replied, "not OK," they were then asked, "Was it just a little bit bad or was it really bad?" Then, children were asked, "If the girl/boy who built that tower earlier came back and saw that it was destroyed, how do you think s/he would feel?" and "If the girl/boy wanted to know who destroyed the tower, what would you say?"

Next, children used the 4-point scaled to rate, "How much do you think that was Robbie the Robot's fault that the tower is destroyed?" and "What about [E2], the other person who was in that other room? How much do you think that was [E2]'s fault that the tower is destroyed?" Children were also asked to rate, "Was Robbie the Robot bad?" and "Was [E2] bad?" Then they rated "Do you think Robbie the Robot should get in trouble for what happened?" and "Do you think [E2] should get in trouble for what happened?" Finally, children rated "Do you think Robbie the Robot knocked down the tower on purpose?" and "Do you think Robbie the Robot knocked down the tower by accident?"

[1] Toward the beginning of data collection for Study 1, in the controlled condition, at times E2 narrated, "Look, I'm making Robbie knock over the tower," but, due to experimenter error, this was not done so consistently with participants. Some time into data collection, it was decided that E2 should never narrate this line, so for the remaining majority of participants, E2 was silent during the destruction of the tower. Unfortunately, no record was kept of which participants heard the version with E2 narrating destruction of the tower, so these two groups could not be compared statistically.

Finally, children were asked to re-rate Robbie the Robot with the same property attribution questions that had been asked previously. The questions were asked in the same order as they had been asked initially. When the interview was finished, children were debriefed and were offered a toy for their participation.

SCORING

Originally, the *thoughts* item and the *intentions* items were to be averaged together to form a composite *agency* score while the *emotions* item and the *sensations* item would form a composite *experience* score, however, initial inspection of the data with factor analysis revealed that that the intentions items loaded uniquely onto a third factor. Thus, the four items in Study 1 were analyzed individually instead. Responses to the two moral concern items were averaged to create a composite *moral concern* score, and responses to the two biological items were averaged to create a composite *biological* attributions score.

Results and Discussion of Study 1

AVERAGE ATTRIBUTIONS

For a general comparison of the robot to other entities and to assess overall effects of age and condition, children's psychological attributions to Robbie the Robot, an average score of *initial* and *final* attributions was computed for each of the four psychological attribution items: intentions, thoughts, emotions, sensations. Internal consistency for these four averaged items, as measured by Cronbach's alpha, was .77. None of the following attributions varied significantly by sex.

Comparison to other Entities

As shown in Figure 4, overall, children attributed more psychological properties (average across the four items) to Robbie the Robot ($M = 1.25$, $SD = 0.84$) than to a computer ($M = 0.44$, $SD = 0.51$), $t(66) = 7.32$, $p < .001$, a tree ($M = 0.74$, $SD = 0.77$), $t(66) = 3.75$, $p < .001$, or a car ($M = 0.32$, $SD = 0.58$), $t(64) = 8.74$, $p < .001$. However, children attributed fewer psychological properties to Robbie the Robot ($M = 1.23$, $SD = 0.93$) than to a human baby ($M = 2.57$, $SD = 0.40$), $t(66) = 12.40$, $p < .001$. Similarly, children were more likely to indicate moral concern for Robbie the Robot ($M = 1.01$, $SD = 1.12$) than for a computer ($M = 0.80$, $SD = 0.84$), $t(66) = 7.45$, $p < .001$, a tree ($M = 1.00$, $SD = 0.99$), $t(66) = 6.37$, $p < .001$, or a car ($M = 0.89$, $SD = 0.85$), $t(65) = 7.43$, $p < .001$, but they were less likely to indicate moral concern for the robot than for a human baby ($M = 2.87$, $SD = 0.41$), $t(66) = 9.76$, $p < .001$.

Children's average attributions of biological properties ("Can [entity] grow?" and "Does [entity] eat?") to Robbie the Robot ($M = 0.26$, $SD = 0.57$) were significantly lower than attributions to a tree ($M = 2.23$, $SD = 0.72$), $t(66) = 17.92$, $p < .001$, or a baby ($M = 2.85$, $SD = 0.40$), $t(66) = 27.50$, $p < .001$. However, children attributed more biological properties to Robbie the Robot than to a computer ($M = 0.04$, $SD = 0.20$), $t(66) = 2.98$, $p = .004$, and children's biological attributions to the robot were not significantly different from their attributions to a car ($M = 0.20$, $SD = 0.52$), $t(64) = 0.91$, $p = .37$. In addition children were more likely to claim that someone built Robbie the Robot ($M = 2.83$, $SD = 0.47$) than that someone built a tree ($M = 0.42$, $SD = 0.91$), $t(66) = 17.55$, $p < .001$, or a baby, ($M = 0.70$, $SD = 1.22$), $t(66) = 13.89$, $p < .001$, but they were just as likely to claim that someone built Robbie the Robot as they were to claim that someone built a computer ($M = 2.87$, $SD = 0.55$) and a car ($M = 2.91$, $SD = 0.42$).

In sum, consistent with previous research (e.g., Jipson & Gelman, 2007; Khan et al., 2011) children clearly viewed Robbie the Robot as a non-biological artifact while simultaneously conceiving of Robbie as partly psychological and worthy of some moral regard. Thus, conceptions of the robot extended across prototypical category boundaries for living and non-living entities, simultaneously exhibiting properties of both of these classes of entity.

Attributions by Condition and Age

Analyses were performed to test the prediction that, compared to children in the controlled condition, children in the autonomous condition would attribute greater levels

of psychological properties, in particular intentions and thoughts, to Robbie the Robot. In addition, it was predicted that, overall, younger children would be more willing to attribute psychological properties to the robot compared to older children.

A 2(condition: autonomous, controlled) × 2 (age group: younger, older) × 4 (attribution-type: intentions, thoughts, emotions, and sensations) mixed analysis of variance (ANOVA) on children's psychological attributions (average of initial and final) to the robot revealed a significant main effect of condition, $F(1, 63) = 13.51$, $p < .001$, $\eta_p^2 = .78$, a significant main effect of age, $F(1, 63) = 25.78$, $p < .001$, $\eta_p^2 = .29$, and a marginally significant effect of attribution-type, $F(2.76, 173.99) = 2.65$, $p = .055$, $\eta_p^2 = .04$. With regard to attribution-type, pairwise comparisons revealed that children attributed more thoughts ($M = 1.44$, $SD = 1.16$) than sensations ($M = 1.05$, $SD = 1.08$), $t(66) = 3.64$, $p = .001$. With regard to condition, children in the autonomous condition ($M = 1.54$, $SD = .87$) attributed significantly greater levels of psychological properties to the robot than children in the controlled condition ($M = 0.97$, $SD = 0.72$), $t(65) = 2.94$, $p = .005$. Regarding age, younger children ($M = 1.64$, $SD = 0.73$) attributed more psychological properties to the robot than did older children ($M = 0.84$, $SD = 0.74$), $t(65) = 4.53$, $p < .001$.

Moral concern for the robot was also assessed with a 2(condition: autonomous, controlled) × 2 (age group: younger, older) ANOVA. Results revealed a significant main effect of condition, $F(1, 63) = 5.70$ $p = .02$, $\eta_p^2 = .08$ and a significant main effect of age, $F(1, 63) = 10.12$, $p = .002$, $\eta_p^2 = .14$. With regard to condition, children in the autonomous condition ($M = 2.02$, $SD = 0.84$) indicated significantly more moral concern

for the robot compared to children in the controlled condition ($M = 1.61$, $SD = 0.78$), $t(65) = 2.08$, $p = .04$. With regard to age, younger children ($M = 2.10$, $SD = 0.79$) indicated more moral concern for the robot than older children did ($M = 1.52$, $SD = 0.78$), $t(65) = 2.97$, $p = .004$.

In sum, as expected, younger children were significantly more likely to anthropomorphize the robot, attributing greater levels of thoughts, intentions, emotions, and sensation to the robot than older children. In contrast, comparison of younger and older children in their attributions of psychological traits to each of the other items (human baby, tree, computer, and car) did not reveal any age differences. Therefore, younger children did not exhibit a general tendency to anthropomorphize more than older children; rather they specifically tended to anthropomorphize the robot more. In addition, manipulation of the robot's apparent autonomy successfully increased children's attributions of psychological properties to the robot, and this was consistent with previous studies using similar manipulations. Whereas previous experiments have shown that increasing cues to self-initiated action increase children's attributions of thoughts (Somanader, Saylor, & Levin, 2011) to a humanoid robot, and emotions and sensations to a robotic dog (Gary & Chernyak, 2013), the current study extends these findings by demonstrating that children are also more willing to attribute intentional behavior (i.e., being able to do things on purpose) to a humanoid robot when it appears to move autonomously. This new finding is of particular interest because of the link between intentional behavior and moral accountability in judgments about human agents (e.g., Cushman, Sheketoff, Wharton, & Carey, 2013). Successful manipulation of the robot's

perceived capacity for intentional behavior allowed for analysis of the potential link between attributions of intention and moral accountability, which will be discussed later.

MORAL ACCOUNTABILITY JUDGMENTS AS A FUNCTION OF CONDITION, AGE, AND SEX

Recall that, after witnessing the destruction of the tower, children were asked rate the severity of the event, and through a structured interview were asked to judge (on a 4-point scale: 0 = "no, not at all"; 1 = "no, not much"; 2 = "yes, a little"; 3 = "yes, a lot") for both Robbie the Robot and E2: Blameworthiness ("How much do you think that was [Robbie's/E2's] fault that the tower is destroyed?"), Naughtiness ("Was [Robbie/E2] bad?"), Punishment ("Do you think [Robbie/E2] should get in trouble for what happened?"). Then children judged for only Robbie: Intent ("Do you think Robbie the Robot knocked down the tower on purpose?"), and Accident ("Do you think Robbie the Robot knocked down the tower by accident?").

Severity Judgments

The mean for severity judgments regarding whether the tower destruction was OK/bad (0 = "OK"; 1 = "a little bad"; 2 = "really bad") was 1.57 (*SD* = 0.61), indicating that, on average, children judged the event as somewhere in between "a little bad" and "really bad." Average severity judgments in the autonomous condition (M = 1.59, SD = 0.61) were not significantly different from judgments in the controlled condition (M = 1.55, SD = 0.62), t(65) = 0.29, p = .78. Children's severity judgments did not vary by age or sex.

Overall Moral Accountability Judgments

It was predicted that children in the autonomous condition would be more likely to view the robot as morally accountable for the tower destruction than children in the controlled condition. In comparison, children were expected to view E2 (the other person in the room) as morally accountable in the controlled condition moreso than in the autonomous condition. Two composite moral accountability scores were created, one for Robbie the Robot and one for E2, by averaging the three moral accountability questions (Blameworthiness: "Was it [Robbie's/E2's] fault that the tower is destroyed?"; Punishment: "Should [Robbie/E2] get in trouble for what happened?"; and Naughtiness: "Was [Robbie/E2] bad?"), respectively. Cronbach's alpha for the three questions about Robbie the Robot's moral accountability was .62. Cronbach's alpha for the three questions about E2's moral accountability was .95.

Children's judgments of Robbie the Robot's moral accountability were submitted to a 2(condition: autonomous, controlled) × 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA. Counter to the original prediction of the study, there was not a significant main effect of condition on judgments of the robot's moral accountability. Overall, judgments in the autonomous condition ($M = 1.75, SD = 0.79$) were not significantly higher than judgments in the controlled condition ($M = 1.60, SD = 0.96$), $t(65) = 0.70, p = .49$. However, the ANOVA revealed a significant condition × sex interaction, $F(1, 59) = 6.30, p = .02, \eta_p^2 = .10$. Boys' judgments of the robot's moral accountability were higher in the autonomous condition ($M = 2.04, SD = 0.52$) than in the controlled condition ($M = 1.41, SD = 1.05$), $t(22.80) = 2.22, p = .04$, Cohen's $d = 0.76$, (although this difference was

non-significant with Bonferroni correction), whereas girls' judgments in the autonomous condition ($M = 1.36$, $SD = 0.93$) were lower than their judgments in the controlled condition ($M = 1.78$, $SD = 0.85$), but this difference was also non-significant, $t(29) = 1.34$, $p = .19$.

A 2(condition: autonomous, controlled) × 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA on children's judgments of E2's moral accountability revealed a significant main effect of condition $F(1, 59) = 34.64$, $p < .001$, $\eta_p^2 = .37$, and a marginally significant main effect of age, $F(1, 59) = 3.05$, $p = .09$, $\eta_p^2 = .05$. Judgments of the E2's moral accountability were significantly higher in the controlled condition ($M = 1.68$, $SD = 1.10$) than in the autonomous condition ($M = 0.35$, $SD = 0.69$), $t(55.70) = 5.94$, $p < .001$, Cohen's $d = 1.45$. Regarding age, an independent-samples t-test did not reveal a significant difference between younger ($M = 0.87$, $SD = 1.06$) and older ($M = 1.18$, $SD = 1.19$) children's judgments of E2's moral accountability, $t(65) = 1.12$, $p = .27$. Therefore, although it appeared that there was not a clear distinction between the autonomous and controlled condition in children's moral accountability judgments about Robbie, the distinction was very clear for judgments about the moral accountability of E2. This suggests that the absence of a condition effect on judgments of Robbie's moral accountability was not due to a failure in understanding the causal link between the controller and activation of the robot's actions. In fact, given the condition effect on judgments about E2, children seemed to have some notion of moral causation through the control mechanism.

Children's judgments of Robbie the Robot's moral accountability were also compared to judgments about E2 by condition. In the autonomous condition, children's moral accountability judgments about Robbie the Robot ($M = 1.75$, $SD = 0.79$) were significantly higher than their moral accountability judgments about E2 ($M = 0.35$, $SD = 0.69$), $t(32) = 7.89$, $p < .001$, Cohen's $d = 1.89$ (see Figure 5). In contrast, in the controlled condition, moral accountability judgments about Robbie the Robot ($M = 1.60$, $SD = 0.96$) were not significantly different than moral accountability judgments about E2 ($M = 1.68$, $SD = 1.10$), $t(33) = 0.37$, $p = .71$. Thus, in the controlled condition, children were equally willing to attribute moral responsibility to Robbie as to E2. In light of this, the absence of a condition effect on judgments of Robbie's moral accountability may have been due to the fact that Robbie was seen as morally responsible in both conditions. Of course, judgments of moral accountability were far from ceiling ($M = 1.60$ out of possible 3), and, anecdotally, several children seemed reluctant to cast a negative moral judgment about an adult stranger (E2), which may have deflated the actual moral accountability scores reported for E2. Nonetheless, it is noteworthy that children's judgments about Robbie and E2 were not orthogonal; children often seemed quite willing to simultaneously assign blame to both, but only when Robbie was explicitly controlled by E2. In fact, 27 out of 34 (79%) of the children in the controlled condition simultaneously viewed both Robbie and E2 as at least partially morally accountable. In contrast, only 9 out of 33 (27%) of children in the autonomous condition viewed both as simultaneously morally accountable.

Blameworthiness

Next, potential effects of condition, age, and sex were assessed for each moral accountability item individually. First, judgments of Robbie's blameworthiness ("Was it was Robbie the Robot's fault that the tower is destroyed?") were explored with a 2 (condition: autonomous, controlled) × 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA. Results did not reveal a main effect of age or any interactions involving age, but there was a significant condition × sex interaction, $F(1, 59) = 8.47$, $p = .005$, $\eta_p^2 = .13$. Follow-up analyses revealed that for the boys, judgments of the robot's blameworthiness were significantly higher in the autonomous condition ($M = 2.53$, $SD = 0.70$) than in the controlled condition ($M = 1.65$, $SD = 1.17$), $t(25.49) = 2.70$, $p = .01$. In contrast, for the girls, judgments of the robot's blameworthiness were higher in the controlled condition ($M = 2.41$, $SD = 0.94$) than in the autonomous condition ($M = 1.79$, $SD = 1.05$), $t(29) = 1.75$, $p = .09$, but this difference was non-significant with Bonferroni correction.

Judgments of the blameworthiness of E2 were also explored with a 2 (condition: autonomous, controlled) x 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA. Results revealed a main effect of condition, $F(1, 59) = 40.99$, $p < .001$, $\eta_p^2 = .41$, and there were no effects or interactions involving age or sex. Children in the controlled condition ($M = 1.94$, $SD = 1.07$) had significantly higher judgments of E2's blameworthiness than children in the autonomous condition ($M = 0.42$, $SD = 0.79$), $t(60.75) = 6.60$, $p < .001$, Cohen's $d = 1.62$.

Naughtiness

Children's judgments of naughtiness, regarding whether Robbie the Robot had been bad were explored with a 2 (condition: autonomous, controlled) x 2 (age group: younger, older) × 2 (sex: girls, boys) ANOVA. Results revealed a main effect of age, $F(1, 59) = 6.79$, $p = .01$, $\eta_p^2 = .10$, and there were no main effects or any interactions involving condition or sex. Older children's judgments ($M = 1.76$, $SD = 1.06$) of Robbie the Robot's naughtiness were significantly higher than younger children's judgments ($M = 1.15$, $SD = 1.21$), $t(65) = 2.19$, $p = .03$, Cohen's $d = 0.54$.

A 2 (condition: autonomous, controlled) x 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA on children's judgments of E2's naughtiness revealed a main effect of condition, $F(1, 59) = 26.61$, $p < .001$, $\eta_p^2 = .31$, and there were no main effects or any interactions involving age or sex. Children's judgments of whether E2 had been bad were significantly higher in the controlled condition ($M = 1.65$, $SD = 1.20$) than in the autonomous condition ($M = 0.36$, $SD = 0.74$), $t(55.21) = 5.27$, $p < .001$, Cohen's $d = 1.29$. Again, this finding highlights the fact that there was a clear distinction between the autonomous and controlled condition in children's moral judgments about E2, even though a similar distinction was not present for judgments about Robbie.

Judgments of Deserved Punishment

A 2 (condition: autonomous, controlled) x 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA on children's judgments of whether Robbie the Robot should get in trouble did not reveal any main effects or interactions. In contrast, a 2 (condition:

autonomous, controlled) x 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA on children's judgments of whether E2 should get in trouble revealed a significant main effect of condition, $F(1, 59) = 23.44$, $p < .001$, $\eta_p^2 = .28$, a marginally significant main effect of age, $F(1, 59) = 3.18$, $p = .08$, $\eta_p^2 = .05$, and a marginally significant main effect of sex, $F(1, 59) = 3.91$, $p = .05$, $\eta_p^2 = .06$. With regard to condition, judgments of whether E2 should get in trouble were significantly higher in the controlled condition ($M = 1.44$, $SD = 1.24$) than in the autonomous condition ($M = 0.27$, $SD = 0.72$), $t(53.35) = 4.75$, $p < .001$, Cohen's $d = 1.15$. With regard to age and sex, follow-up t-tests did not reveal any significant differences, however younger children ($M = 0.71$, $SD = 1.09$) tended to be slightly less likely than older children ($M = 1.03$, $SD = 1.24$) to claim that Robbie deserved punishment, $t(65) = 1.14$, $p = .26$.

Intent Judgments

A 2 (condition: autonomous, controlled) x 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA on children's judgments of intent, that is, whether Robbie the Robot knocked down the tower on purpose, revealed a main effect of condition, $F(1, 41) = 4.66$, $p = .04$, $\eta_p^2 = .10$, and there were no main effects or any interactions involving age or sex. Children's judgments of intent were significantly higher in the autonomous condition ($M = 1.63$, $SD = 1.41$) than in the controlled condition ($M = 0.84$, $SD = 1.21$), $t(47) = 2.09$, $p = .04$, Cohen's $d = 0.60$. A 2 (condition: autonomous, controlled) x 2 (age group: younger, older) × 2(sex: girls, boys) ANOVA on children's judgments of whether Robbie the Robot knocked down the tower by accident revealed a main effect of age, $F(1, 41) =$

10.63, $p = .002$, $\eta_p^2 = .21$, and there were no main effects or interactions involving condition or sex. Younger children's judgments ($M = 2.30$, $SD = 1.10$) of whether Robbie the Robot knocked down the tower by accident were significantly higher than older children's judgments ($M = 1.23$, $SD = 1.38$), $t(39.88) = 2.95$, $p = .005$, Cohen's $d = 0.86$. This finding, along with the finding that 5-year-olds were less likely to claim that Robbie had been bad, suggests that the younger children were more willing to forgive Robbie even if, compared to older children, they were just as likely to recognize that the robot had been at fault.

MORAL JUDGMENTS AS A FUNCTION OF PSYCHOLOGICAL ATTRIBUTIONS

A series of multiple regression analyses was used to determine whether children's judgments of the severity of the destructive behavior and of Robbie the Robot's moral accountability varied as a function of children's psychological attributions to the robot. For these analyses, composite scores were created for each psychological attribution item (intentions, thoughts, emotions, sensations) by averaging initial and final responses for each one. This was meant to be an index of how children conceived of the robot overall throughout the duration of the experiment.

Severity Judgments

Multiple regression analysis was used to determine whether children's severity judgments varied as a function of intentions, thoughts, emotions, and sensations. Results

indicated that none of the psychological attributions significantly predicted severity judgments, $F(4, 62) = 1.12, p = .36$.

Moral Accountability Judgments

It was hypothesized that children's psychological attributions, in particular attributions of agency, would predict their willingness to hold Robbie the Robot morally accountable for having destroyed the tower. Multiple regression analysis was used to determine whether Moral Accountability varied as a function of intentions, thoughts, emotions, and sensations.

Results indicated that attribution of intentions ($\beta = .24, p = .03$) significantly predicted moral accountability judgments. A forward selection stepwise regression resulted in a model with only intentions as a significant predictor ($\beta = .22, p = .049$), $R^2 = .06, F(1, 65) = 4.04, p = .049$. With more attributions of intentions, children tended to be more willing to hold the robot morally accountable. Separate regression analyses for the autonomous condition and the controlled condition were performed as well. For the autonomous condition, no predictors emerged as significant, but for the controlled condition, attribution of intentions ($\beta = .40, p = .049$) significantly predicted moral accountability judgments, $R^2 = .12, F(1, 32) = 4.17, p = .049$. Thus, attributions of intentions to Robbie the Robot ("Can Robbie do things on purpose?") positively predicted the degree to which children held the robot morally accountable for destroying the tower, specifically in the controlled condition.

A multiple regression analysis on children's moral accountability judgments about E2 as a function of psychological attributions to the robot revealed that attributions of thought ("Can Robbie think?") to Robbie the Robot ($\beta = -.53, p = .001$) significantly predicted children's judgments. A forward-selection stepwise regression resulted in a model with only thoughts as a significant predictor ($\beta = -.48, p < .001$), $R^2 = .24$, $F(1, 65) = 20.42$, $p < .001$. Separate regression analyses by condition indicated no significant predictors of E2's moral accountability in the autonomous condition, however, in the controlled condition, a forward-selection stepwise regression resulted in a model with only thoughts as a significant predictor ($\beta = -.49, p = .003$), $R^2 = .24$, $F(1, 32) = 10.09$, $p = .003$. Therefore, the more thoughts children attributed to Robbie, the less they were inclined to hold E2 morally accountable, specifically in the controlled condition (the only condition in which children were likely to assign any blame to E2 in the first place).

Intent Judgments

Multiple regression analysis was used to determine whether children's judgments of Robbie the Robot's Intent, that is, whether the robot knocked down the tower on purpose, varied as a function of each of the four psychological attributions to the robot. Results indicated that attribution of intentions was a significant predictor ($\beta = .68, p < .001$), $F(4, 44) = 4.04$, $p = .007$. A forward-selection stepwise regression resulted in a model with only intentions as a significant predictor ($\beta = .65, p < .001$), $R^2 = .23$, $F(1, 47) = 14.21$, $p < .001$. Separate regression analyses by condition indicated no significant predictors of intent in the controlled condition, however, in the autonomous condition, a

forward-selection stepwise regression resulted in a model with only intentions as a significant predictor ($\beta = .68, p = .01$), of judgments of Robbie's intent in knocking over the tower, $R^2 = .25, F(1, 22) = 7.48, p = .01$. Thus, the more children attributed to Robbie the ability to do things on purpose, the more likely they were to see the act of destruction as intentional, specifically in the autonomous condition.

A multiple regression analysis on children's judgments of whether the robot knocked down the tower by accident indicated no significant predictors. However, separate analyses by condition revealed that, in the controlled condition alone, attribution of emotions ($\beta = .62, p = .004$) significantly predicted judgments of whether the robot knocked over the tower by accident, $R^2 = .56, F(1, 23) = 10.43, p = .004$. Hence, the more that children attributed to Robbie the Robot the capacity to feel things like happy or sad, the more children tended to claim that Robbie destroyed the tower by accident, in the controlled condition specifically. Attributions of emotions and accident judgments were significantly correlated ($r = .56, p = .004$) and remained significantly correlated when controlling for age, ($r = .44, p = .03$).

Mediation Between Condition and Moral Accountability

Given the significant effect of condition on children's attributions of intentions and the positive association between attributions of intentions and judgments of moral accountability ($\beta = .22, p = .049$), it was hypothesized that the relationship between condition and moral accountability judgments may have been mediated by attribution of intentions. Note that such a mediation is still possible even if the total effect (in this case

the effect of condition on moral accountability judgments) is not statistically significant (Shrout & Bolger, 2002). A mediation analysis was performed following the Preacher and Hayes Multiple Mediation procedure for estimating indirect effects (Hayes & Preacher, 2014). This procedure was chosen over more traditional mediation tests (e.g., Baron & Kenny, 1986) because it is more robust in detecting indirect effects, especially with small samples, due to its use of bootstrapping method.

Results indicated that the relationship between condition (dummy coded: controlled = 0; autonomous = 1) and moral accountability judgments (composite score) was mediated by children's attributions of intentions to Robbie the Robot (average of initial and final). As shown in Figure 6, the pathway between condition and attribution of intentions, indicated by the standardized regression coefficient, was significant, and so was the pathway between attributions of intentions and moral accountability judgments. There was a standardized indirect effect of (.62)(.21) = .13. To test the significance of this indirect effect, the 95% confidence interval was determined with 1,000 bootstrapping resamples. Results yielded a bootstrapped unstandardized indirect effect of .13, and the 95% confidence interval ranged from .003, 0.381, indicating that the indirect effect was statistically significant. The reduction in the model was 86.67%, indicating a large mediation of the effect of condition on moral accountability judgments through the attribution of intentions.

In light of this result, the absence of a significant difference in moral accountability judgments about Robbie the Robot in the autonomous versus the controlled condition can be explained by the fact that the effect was largely mediated by

attributions of intentions to the Robot. Hence, increasing cues to the robot's autonomy only affected moral accountability judgments through an increase in children's attributions of intentions.

CHANGES IN PSYCHOLOGICAL ATTRIBUTIONS AFTER DESTRUCTIVE BEHAVIOR

Potential changes in children's psychological attributions were examined by comparing initial and final attributions. A 2 (condition: autonomous, controlled) × 4 (attribution-type: intentions, thoughts, emotions, sensations) × 2 (time: initial, final) mixed-design ANOVA on children's attributions revealed a significant main effect of condition, $F(1, 65) = 8.63$, $p = .005$, $\eta_p^2 = .12$, and a significant interaction of attribution-type and time, $F(3, 195) = 5.18$, $p = .002$, $\eta_p^2 = .07$. Follow-up comparisons revealed that attributions of intentions significantly increased from initial ($M = 0.94$, $SD = 1.20$) to final ($M = 1.46$, $SD = 1.23$), $t(66) = 2.97$, $p = .004$, Cohen's $d = .43$. Attributions of thoughts remained similar from initial ($M = 1.45$, $SD = 1.32$) to final ($M = 1.43$, $SD = 1.29$), $t(66) = 0.10$, $p = .92$. Attributions of emotions decreased slightly from initial ($M = 1.37$, $SD = 1.32$) to final ($M = 1.25$, $SD = 1.16$), and the same was the case for sensations from initial ($M = 1.16$, $SD = 1.25$) to final ($M = 0.94$, $SD = 1.20$), but neither of these comparisons were statistically significant, $t(66) = 0.78$, $p = .44$, and $t(66) = 1.56$, $p = .13$, respectively. Figure 7 illustrates psychological attributions by attribution type, condition, and time.

Although the ANOVA did not reveal a significant interaction of condition with attribution type and time, planned comparisons were performed to examine changes in

psychological attributions by condition. For the autonomous condition, a 4 (attribution-type: intentions, thoughts, emotions, sensations) × 2 (time: initial, final) repeated-measures ANOVA on children's attributions revealed a significant interaction of attribution-type and time, $F(3, 96) = 3.49$, $p = .02$, $\eta_p^2 = .10$. Follow-up comparisons revealed that attributions of intentions significantly increased from initial ($M = 1.18$, $SD = 1.38$) to final ($M = 1.85$, $SD = 1.12$), $t(32) = 2.81$, $p = .008$, Cohen's $d = .13$. Attributions of thoughts increased slightly but non-significantly from initial ($M = 1.73$, $SD = 1.33$) to final ($M = 1.94$, $SD = 1.22$), $t(33) = 0.96$, $p = .34$. Attributions of emotions remained the same from initial ($M = 1.48$, $SD = 1.33$) to final ($M = 1.48$, $SD = 1.35$), and attributions of sensations decreased slightly from initial ($M = 1.42$, $SD = 1.32$) to final ($M = 1.24$, $SD = 1.39$), but this change was not significant, $t(32) = 0.83$, $p = .41$. For the controlled condition, a 4 (attribution-type: intentions, thoughts, emotions, sensations) × 2 (time: initial, final) repeated-measures ANOVA on children's attributions did not reveal any significant main effects or interaction.

Therefore, in the autonomous condition only, there was a significant increase in attributions of intentions (see Figure 7). This shows that children were likely to update their ideas about whether the robot could do things on purpose and that they were sensitive to cues suggesting goal-directed behavior. In addition, it is possible that the moral nature of the action may have contributed to children's shift in attributions. This is consistent with a previous study showing that when a humanoid robot was observed "cheating" during a game of rock-paper-scissors against an adult, participants playing the game tended to attribute agentive mental states to the robot (Short, Hart, Vu &

Scassellati, 2010). As the authors point out, the act of cheating suggests a goal-directed behavior with a desire to win something. Admittedly, judgments of the robot's intentionality by adults in that study, and by children in the current study, could have similarly been influenced by witnessing an apparently goal-directed *neutral* behavior. Nonetheless, morally *bad* goal-directed behaviors may be especially noteworthy to people, and some studies suggest that judgments of an actor's intentions can change depending on the moral outcome of an action (Knobe, 2003). Even children as young as 5 years old may be biased to claim that an action was intentional if the outcome of that action is negative relative to when the outcome is positive or neutral (Leslie, Knobe, Cohen, 2006). With regard to the current study, whether judgments of Robbie the Robot's perceived intentionality were affected by the *moral* nature of the event is still an open question. This issue was further explored in Study 2.

PARENT QUESTIONNAIRE (PQ)

Finally, analyses were conducted to determine whether psychological attributions and moral accountability judgments varied as a function of children's prior exposure to robots based on responses to the parent questionnaire: whether children had seen a robot in person before (dummy coded: no = 0, yes = 1), children's experience with robotic toys, frequency viewing media with robot characters, whether parents conversed with their children about how robots work, and children's experience with animals such as dogs and cats.

PQ and Psychological Attributions

Multiple regression analysis was used to determine whether children's psychological attributions to Robbie the Robot (collapsed across condition) varied as a function of each of the five items from the parent questionnaire. A forward-selection stepwise regression on children's attributions of thoughts (average of initial and final) did not reveal any significant predictors. A forward-selection stepwise regression on children's attributions of intentions (average of initial and final) did not reveal any significant predictors either. A forward-selection stepwise regression on children's attributions of emotions (average of initial and final) resulted in a model with children's prior exposure to a robot that they had seen in person ($\beta = -.72, p = .02$) as the only significant predictor, $R^2 = .09, F(1, 64) = 6.29, p = .02$. Children who had seen a robot before were less inclined to attribute emotions to Robbie. A forward-selection stepwise regression on children's attributions of sensations (average of initial and final) did not reveal any significant predictors.

Potential relations between the items from the parent questionnaire and psychological attributions were also assessed by condition. Regression analysis on children's attributions of thoughts (average of initial and final) did not reveal any significant predictors in the autonomous condition or in the controlled condition. A forward-selection stepwise regression on children's attributions of intentions (average of initial and final) in the autonomous condition did not reveal any significant predictors either. However, in the controlled condition, a forward-selection stepwise regression on children's attributions of intentions (average of initial and final) resulted in a model with

children's experience with animals ($\beta = -.24$, $p = .04$) as the only significant predictor, $R^2 = .14$, $F(1, 31) = 4.82$, $p = .04$. The more experience children had interacting with animals such as dogs or cats, the less they tended to attribute intentions to Robbie the Robot in the controlled condition specifically. With regard to emotions (average of initial and final), a forward-selection stepwise regression on children's attributions in the autonomous condition resulted in a model with child-parent conversations about how robots function ($\beta = -.43$, $p = .04$) as the only significant predictor, $R^2 = .13$, $F(1, 31) = 4.63$, $p = .04$. The more experience children had talking to their parents about how robots function, the less they tended to attribute emotions to Robbie the Robot in the autonomous condition specifically. In the controlled condition forward-selection stepwise regression on children's attributions of emotions (average of initial and final) resulted in a model with whether children had seen a robot in person before ($\beta = -.90$, $p = .04$) as the only significant predictor, $R^2 = .14$, $F(1, 31) = 4.86$, $p = .04$. Children who had seen a robot before were less inclined to attribute emotions to Robbie specifically in the controlled condition. Finally, with regard to sensations (average of initial and final), no significant predictors emerged in the autonomous condition or in the controlled condition.

In sum, with exposure to robots, having previously seen one in person, and with increased knowledge about their functions through parent conversations, children were less willing to attribute emotions to Robbie the Robot. Interestingly, with more exposure to animals, children were also less inclined to attribute intentions to the robot.

PQ and Moral Accountability

When responses were collapsed across condition, no items from the parent questionnaire emerged as significant predictors of children's moral judgments about Robbie the Robot. However, when conditions were examined separately, a few relationships emerged between items from the parent questionnaire and scores for the moral accountability questions. For judgments of Robbie's naughtiness (whether Robbie had been bad) in the controlled condition, a forward-selection stepwise regression resulted in a model with parent conversations about robot functioning ($\beta = .41, p = .04$) as the only significant predictor, $R^2 = .13, F(1, 31) = 4.64, p = .04$. Counter to the predictions, the more that parents had spoken to children about robots, the more willing children were to claim that Robbie had been bad, in the controlled condition. Regarding judgments of whether Robbie knocked down the tower on purpose, in the autonomous condition, a forward-selection stepwise regression resulted in a model with parent conversations about robot functioning ($\beta = -.69, p = .009$) and previous interaction with animals ($\beta = .63, p = .05$) as the only significant predictors, $R^2 = .32, F(2, 21) = 5.01, p = .02$. The less experience children had talking to their parents about how robots work and the more experience children had interacting with animals, the more that children were inclined to claim that Robbie had knocked over the tower on purpose in the autonomous condition. No predictors emerged for judgments about whether the tower destruction had been an accident, whether Robbie was blameworthy, or whether Robbie deserved to get in trouble.

SUMMARY

Findings from Study 1, along with previous studies (Gary, 2014; Gary & Chernyak 2013; Somanader, Saylor, & Levin, 2011) demonstrate that young children's attributions of psychological properties and moral status to a humanoid robot are affected by the perceived source of the robot's action, that is, whether it appears to move of its own accord. The current study is the first to demonstrate that young children's judgments of a robot's agency, specifically the capacity for intentional behavior, are also significantly affected by such cues and that, via the attribution of intentions, children are willing to hold a robot morally accountable for causing harm. In addition, the results suggest that seeing a robot engage in a morally charged action may increase children's attributions of psychological agency but not experience to the robot. In particular, children's judgments of whether Robbie the Robot could do things on purpose increased after they had seen Robbie cause harm, specifically in the autonomous condition in which there was not a person apparently controlling the robot. In contrast, judgments about Robbie's capacity to feel and think were relatively unaffected by this event.

Study 2 was meant to further explore the idea that making a robot appear to be a moral agent would specifically affect children's attributions of psychological agency but not experience to it (Gray & Wegner, 2009). This hypothesis was explored mainly by including a much wider range of psychological attribution questions pertaining to both agency and experience. In addition, since the main interest was to explore potential changes in these attributions in an autonomous condition, the procedure was simplified to only include one experimenter, the interviewer. Since there were no concerns about

paralleling the setup between two conditions (as in Study 1), the absence of anyone else in the room was leveraged to create a more convincing illusion of the robot's autonomy, which was meant to elicit a richer response from children. An additional methodological change was made for practical purposes but also may have served to increase the robot's apparent autonomy: children witnessed the morally charged event (tower destruction) in a video that was purported to be a live feed from the room next door. As in Study 1, children's attributions were elicited before and after the event, which allowed for an analysis of potential changes in children's conception of the robot's psychological status.

Method of Study 2

Participants

Participants were 21 5-year-olds (mean age = 5;3, range = 5;0-5;10; 15 boys and 6 girls), and 23 7-year-olds (mean age = 7;6, range = 7;2-8;0; 15 boys and 8 girls). Participant ethnicity was 66% Caucasian, 18% Hispanic or Latino, 9% Asian or Pacific Islander, 9% more than one race, and 2% African American. Participants were recruited from the database at the UT Children's Research Lab.

Materials

The same Nao robot used in Study 1 was used in Study 2. The main difference in the second study was the inclusion of a more diverse range of psychological attribution questions, partly modeled from items used in previous studies tapping into multiple dimensions of mind perception (e.g., Gray, et al., 2007). Eight questions pertained to psychological agency and thoughtfulness: "Can Robbie control his own actions?"; "Can Robbie make plans?"; "Can Robbie make decisions?"; "Can Robbie do things on purpose?"; "Can Robbie remember things?"; "Can Robbie think?"; "Can Robbie have ideas?"; "Does Robbie have a mind?" The other eight items pertained to psychological experience, in particular negative and positive emotional experience: "Can Robbie feel upset"; "Can Robbie feel scared?"; "Can Robbie feel hurt?"; "Can Robbie feel sad?"; Can Robbie feel loved?"; "Can Robbie feel excited?"; "Can Robbie feel good?"; "Can Robbie feel happy?" Each question was printed out and laminated to create a set of 16 cards. The same picture of Robbie the Robot used in the first study was shown to children

during the psychological attributions interview. The same 4-point scale was also used: large thumbs-down ("no, not all), a small thumbs-down ("no, not much") a small thumbs-up ("yes, a little"), and large thumbs-up ("yes, a lot"). No other entities were included in the interview.

In contrast to Study 1, children witnessed the tower destruction event through a purportedly live feed from a camera in the room next door. In reality, children were shown a prerecorded video taken from an elevated position in one corner of the room. Within the frame of the video was most of the room with Robbie the Robot initially sitting on the right-hand side of the video and the block tower on the left. For the first 10 seconds of the video, nothing seemed to be happening. Then, suddenly, Robbie was seen standing up, turning and walking in the general direction of the block tower. The robot could be seen coming to a stop right next to the tower, and, after a brief pause, proceeding to destroy it with a punch.

PROCEDURE

Study 2 procedure was very similar to Study 1 with a few noteworthy differences. As in Study 1, children were first escorted into the room to see Robbie the Robot, who was sitting in the corner of the room across from the participant. Unlike in Study 1, in Study 2 there was only one experimenter, the interviewer, and that experimenter also inconspicuously initiated the robot's actions by pressing a button on a hand-held touch device hidden behind a clipboard while children were looking at the robot. As in Study 1, the robot proceeded to stand up, stretch, walk in the direction of the child, wave, and sit

down. The child was then told, "Let's go back to the other room now," and as s/he stood up from their chair, "Oh hey, did you notice that block tower over there? There was a [girl/boy, matched to participant gender] who built that tower earlier today. [S/he] spent a lot of time and was really proud of it, so we're saving it for [her/him] so [s/he] can come back later to show some friends. Do you like it?" Then, children were escorted back to the interview room, and doors were kept closed to create a sense of partition from the robot during questioning. In contrast to Study 1, no other entities were included in the interview.

In the next part, children were introduced to the 4-point scale they were to use to rate their answers to questions. After the three warm up questions, children were asked to rate how much they *liked* Robbie, whether they thought Robbie was *scary* at all, and whether they thought Robbie was *cute*. The experimenter then placed the picture of Robbie in front of the child, the 16 psychological attribution items were shuffled, and children were asked to rate their answer to each question. These answers constituted the *initial* psychological attributions.

During the next part, rather than return to the other room as was done in Study 1, children were instead told, "OK, that's all the questions I have for now. Now I want to show you something. There's actually a camera hidden in the other room, and we can watch to see what Robbie is doing right now. Want to see?" The experimenter then turned on a computer screen directly across from the child to reveal the "surveillance" image of the room next door with Robbie sitting on the floor, and with an inconspicuous press of the spacebar, the experimenter started the secretly prerecorded video. During the

first 10 seconds of inactivity, the experimenter explained, “See? There’s the room. And there’s Robbie… Looks like Robbie’s just sitting there…” The experimenter looked down at the clipboard to pretend like nothing else was happening until the robot began to move, at which time the child typically called the attention of the experimenter, and both watched as Robbie walked over and destroyed the tower. Immediately after, the experimenter turned off the screen, turned to the child, and asked, “What just happened?”

The next part consisted of the moral accountability interview, which was very similar to Study 1 but with a few changes. First, children were asked to rate the *severity* of the event: “Do you think what just happened with those blocks was OK or NOT OK? / Can you show me on the scale?” After a few check questions regarding the implications of the tower destruction on the anonymous builder child’s feelings, children were asked to rate the robot’s *blameworthiness*: “How much do you think that was Robbie the Robot’s fault that the tower is destroyed?”; *intent*: “Do you think Robbie knocked down the tower on purpose?”; *accidental status*: “Do you think Robbie the Robot knocked down the tower by accident?”; *deserved punishment*: “Do you think Robbie should get in trouble for what happened?” and “Should Robbie be put in time-out?”; and *naughtiness*: “Was Robbie bad?” Finally, children were asked to explain why they thought Robbie had knocked over the tower.

Children were then asked to rerate Robbie the Robot on the 16 psychological attribution questions (in the same shuffled order as before), and this constituted children’s *final* attributions. Then, children were asked a series of open-ended questions about how they thought robots worked, and from where they thought the source

of a robot's control typically originates. At the end, children were debriefed about the about the setup of the experiment. They were told that there wasn't actually a child who built the tower so no one was going to be sad. Then they were shown that the video they had seen was in reality prerecorded, and they were told that the experimenter had actually controlled the robot. Children were then given the opportunity to play with Robbie by controlling the robot themselves. Finally, they were offered a small toy prize, and they were thanked for their time.

Results and Discussion of Study 2

Initial Psychological Attributions

The 16 initial psychological attribution questions were submitted to a Principle-components factor analysis, using maximum likelihood extraction with an oblimin rotation (delta = 0). Three strong factors were indicated (Factor 1 eigenvalue = 7.06, 44.14% variance explained; Factor 2 eigenvalue = 2.15, 13.41% variance explained; Factor 3 eigenvalue = 1.28, 8.00% variance explained). See Table 1 for factor loadings. Making plans, remembering, thinking, and having a mind loaded strongly onto the first factor, so these four items (Chronbach's alpha = .82) were averaged together to create a *cognition* composite score. Feeling loved, excited, good, and happy loaded strongly onto the second factor; these four items (Chronbach's alpha = .92) were averaged together as a *positive emotions* score. Feeling upset, scared, hurt, and sad loaded strongly onto the third factor; these four items (Chronbach's alpha = .79) were averaged together as a *negative emotions* score. A fourth factor with an eigenvalue of 1.14, explained 7.14% of the variance, but notably only one of the 16 items, the *intentions* item ("Can Robbie do things on purpose?") loaded strongly onto this fourth factor. Thus, the intentions item was considered individually in subsequent analyses. The decisions item and the ideas item had moderate cross-loadings with more than one factor. In addition, the control item loaded strongly onto the second factor along with positive emotions. For these reasons, decisions, ideas, and control were not included in the cognition composite score.

Recall that children's ratings on the scale were: 0 = "no, not at all"; 1 = "no, not much"; 2 = "yes, a little"; 3 = "yes, a lot." Mean initial attribution of negative emotions was 1.46 ($SD = 0.91$); mean attribution of positive emotion was 2.48 ($SD = 0.79$); mean attribution of cognition was 2.10 ($SD = 0.86$). A 3 (attribution-type: negative emotion, positive emotion, cognition) × 2 (age group: younger, older) × 2(sex: girls, boys) repeated-measures ANOVA on children's initial judgments revealed a significant main effect of attribution-type, $F(2, 78) = 22.49$, $p < .001$, $\eta_p^2 = .37$. There were no main effects or interactions involving age or sex. Follow-up analyses revealed that children attributed significantly more positive emotion ($M = 2.48$, $SD = 0.79$) than negative emotion ($M = 1.46$, $SD = 0.92$), $t(42) = 7.48$, $p < .001$, Cohen's d =1.19, and cognition ($M = 2.10$, $SD = .86$), to Robbie the Robot, $t(42) = 3.48$, $p = .001$, Cohen's d =0.46. Attributions of cognition were also significantly higher than attributions of negative emotion, $t(43) = 4.02$, $p < .001$, Cohen's d =0.72.

The absence of an effect of age on psychological attributions was unexpected, given the significant difference between 5- and 7-year-olds in Study 1. In an attempt to explore possible reasons for this inconsistency, initial psychological attributions in the autonomous condition in Study 1 were compared to initial attributions in Study 2 (all were autonomous) separately by age group. Comparing the average attributions across the 12 items from Study 2 to the average attributions across the four items from Study 1 revealed that the younger children's attributions were similar in both studies (Study 1: $M = 1.84$, $SD = 1.00$; Study 2: $M = 2.09$, $SD = 0.54$), $t(34) = 0.89$, $p = .31$, whereas the older children's attributions were significantly higher in Study 2 ($M = 1.96$, $SD = 0.77$)

compared to Study 1 ($M = 1.09$, $SD = 0.95$), $t(38) = 3.22$, $p = .003$. This suggests that the methodological changes made in Study 2 had a significant impact on the older children's psychological attributions to Robbie. That is, for the 7-year-olds but not the 5-year-olds, it seemed that not having a second experimenter sitting in the room and narrating the robot's actions resulted in making the robot appear even more autonomous and agentive. It is interesting that this made a difference only for the older children. One possibility is that in Study 2, older children were more sensitive to the absence of anyone else in the room, and conversely, they were also more sensitive to the presence of E2 in Study 1. Indeed, in Study 2, upon seeing the robot move for the first time, several of the older children spontaneously asked the experimenter about how the robot was moving by itself.

MORAL ACCOUNTABILITY JUDGMENTS

Severity Judgments

Children's severity judgments regarding whether the tower destruction was OK ($M = 0.64$, $SD = 0.89$), indicated that, on average, children judged the event as somewhere in between "not all OK" and "not much OK." Multiple linear regression analysis was performed to assess whether judgments of the severity of the tower destruction varied as a function of age (treated as a continuous variable), sex (dummy coded), and both initial *and* final attributions of positive emotion, negative emotion, cognition, and intentions. No significant predictors emerged.

Overall Moral Accountability

Moral accountability was examined with a composite score, an average of children's responses to four moral accountability questions: blameworthiness (was it Robbie's fault?), punishment (should Robbie get in trouble?), time-out (should Robbie be put in time-out?), and naughtiness (was Robbie bad?). Cronbach's alpha for these four items was .76.

Multiple linear regression analysis was performed to assess whether composite moral accountability scores varied as a function of age (treated as a continuous variable), sex (dummy coded), and initial *and* final attributions of: positive emotion, negative emotion, cognition, and intentions. A forward-selection stepwise regression resulted in a model with intentions ($\beta = .32, p = .003$) and age ($\beta = .29, p = .01$) as the only significant predictors, $R^2 = .31, F(2, 37) = 8.22, p = .001$. Thus, as children attributed more intention to Robbie, that is, the capacity to do things on purpose, they were also more willing to judge Robbie the Robot as morally accountable for having destroyed the tower. In addition, the older the child the more s/he was willing to judge Robbie the Robot as morally accountable.

Blameworthiness

Multiple linear regression analysis was performed to assess whether judgments of blameworthiness ("Was it Robbie's fault that the tower was destroyed?"), varied as a function of age (treated as a continuous variable), sex (dummy coded), and *initial* attributions of positive emotion, negative emotion, cognition, and intentions. None of the

predictors emerged as significant when they were all entered together into the model, however a forward-selection stepwise regression resulted in a model with intentions (β = .32, p = .009) as the only significant predictor, R^2 = .16, $F(1, 41) = 7.60$, p = .009.

Multiple linear regression analysis was also performed to assess whether blameworthiness varied as a function of age (treated as a continuous variable), sex (dummy coded), and *final* attributions of positive emotion, negative emotion, cognition, and intentions. A forward-selection stepwise regression resulted in a model with intentions (β = .31, p = .02) as the only significant predictor, R^2 = .13, $F(1, 39) = 5.97$, p = .02.

Intent Judgments

Multiple linear regression analysis was performed to assess whether judgments of intent (average of "Did Robbie knock over the tower on purpose?" and reversed scored, "Did Robbie knock over the tower by accident?"; Cronbach's alpha = .77) varied as a function of age (treated as a continuous variable), sex (dummy coded), and *initial* attributions of positive emotion, negative emotion, cognition, and intentions. Cognition (β = .29, p = .03) was a significant predictor and intentions (β = .20, p = .08) was marginally significant. A forward-selection stepwise regression resulted in a model with intentions (β = .37, p = .001) and cognition (β = .39, p = .02) as the only significant predictors, R^2 = .32, $F(2, 40) = 9.54$, $p < .001$.

Multiple linear regression analysis was also performed to assess whether intent judgments varied as a function of age (treated as a continuous variable), sex (dummy

coded), and *final* attributions of positive emotion, negative emotion, cognition, and intentions. A forward-selection stepwise regression resulted in a model with intentions ($\beta = .40, p = .004$) as the only significant predictor, $R^2 = .19, F(1, 39) = 9.31, p = .004$.

Deservedness of Punishment

Multiple linear regression analysis was performed to assess if judgments of whether Robbie should be punished (average of: "Do you think Robbie should get in trouble?" and "Should Robbie be put in time-out?"; Cronbach's alpha = .86) varied as a function of age (treated as a continuous variable), sex (dummy coded), and *initial* attributions of positive emotion, negative emotion, cognition, and intentions. A forward-selection stepwise regression resulted in a model with age ($\beta = .34, p = .02$) as the only significant predictor, $R^2 = .12, F(1, 41) = 5.47, p = .02$.

Multiple linear regression analysis was also performed to assess whether deservedness of punishment varied as a function of age (treated as a continuous variable), sex (dummy coded), and *final* attributions of positive emotion, negative emotion, cognition, and intentions. A forward-selection stepwise regression resulted in a model with age ($\beta = .36, p = .02$) and intentions ($\beta = .32, p = .02$) as the only significant predictors, $R^2 = .23, F(2, 38) = 5.71, p = .007$.

Naughtiness

Multiple linear regression analysis was performed to assess whether judgments of naughtiness ("Was Robbie bad?") varied as a function of age (treated as a continuous

variable), sex (dummy coded), and *initial* attributions of positive emotion, negative emotion, cognition, and intentions. A forward-selection stepwise regression resulted in a model with intentions ($\beta = .40, p = .006$) as the only significant predictor, $R^2 = .17, F(1, 41) = 8.51, p = .006$.

Multiple linear regression analysis was also performed to assess whether judgments of naughtiness varied as a function of age (treated as a continuous variable), sex (dummy coded), and *final* attributions of positive emotion, negative emotion, cognition, and intentions. A forward-selection stepwise regression resulted in a model with intentions ($\beta = .32, p = .05$) as the only predictor, $R^2 = .10, F(1, 39) = 4.27, p = .05$.

Thus, consistently, across all moral accountability judgments, attribution of intentions ("Can Robbie do things on purpose?") was arguably the best predictor of whether children were willing to hold Robbie responsible for destroying the tower. This result is consistent with findings from Study 1, and it is consistent with previous research that demonstrates that adult's attributions of psychological agency positively predict moral responsibility placed on a range of nonhuman entities (Gray, et al., 2007).

CHANGES IN PSYCHOLOGICAL ATTRIBUTIONS FOLLOWING DESTRUCTIVE ACT

Potential changes in children's psychological attributions following the tower destruction event were assessed by comparing initial and final attributions. A 4 (attribution-type: negative emotion, positive emotion, cognition, intention) × 2 (time: initial, final) × 2 (age group: younger, older) × 2(sex: girls, boys) repeated-measures ANOVA on children's attributions revealed a significant main effect of attribution-type,

$F(3, 108) = 10.97$ $p < .001$, $\eta_p^2 = .23$, which was qualified by a three-way interaction of attribution-type, time, and age, $F(3, 108) = 6.36$ $p = .001$, $\eta_p^2 = .15$, and a four-way interaction of attribution-type, time, age, and sex, $F(3, 108) = 3.74$ $p = .01$, $\eta_p^2 = .09$.

To explore this interaction, separate analyses were performed to examine effects for each attribution-type individually. A 2 (time: initial, final) × 2 (age group: younger, older) × 2(sex: girls, boys) repeated-measures ANOVA on children's attributions of negative emotions did not reveal any main effects or interactions. Similarly, a 2 (time: initial, final) × 2 (age group: younger, older) × 2(sex: girls, boys) repeated-measures ANOVA on children's attributions of positive emotion did not reveal any main effects or interactions.

In contrast, a 2 (time: initial, final) × 2 (age group: younger, older) × 2(sex: girls, boys) repeated-measures ANOVA on children's attributions of cognition revealed a marginally significant interaction of time and sex, $F(1, 40) = 4.20$, $p = .05$, $\eta_p^2 = .10$. For girls, initial ($M = 1.84$, $SD = 0.92$) and final ($M = 1.98$, $SD = 0.87$) attributions were not significantly different. For boys, attributions of cognition decreased marginally significantly from initial ($M = 2.23$, $SD = 0.82$) to final ($M = 2.00$, $SD = 0.87$), $t(29) = 1.91$, $p = .07$, Cohen's $d = 0.27$, however with Bonferroni correction, this difference was non-significant.

A 2 (time: initial, final) × 2 (age group: younger, older) × 2(sex: girls, boys) repeated-measures ANOVA on children's attributions of intentions revealed a marginally significant main effect of sex, $F(1, 39) = 1.95$, $p = .07$, $\eta_p^2 = .08$, and a significant interaction of time and age, $F(1, 39) = 10.46$, $p = .002$, $\eta_p^2 = .21$. With regard to the main

effect of sex, the difference in attributions of intentions between boys ($M = 2.03$, $SD = 1.07$) and girls ($M = 1.46$, $SD = 0.97$) was not significant. With regard to the interaction of age and time, younger children's attributions of intentions ("Can Robbie do things on purpose?") significantly increased from initial ($M = 1.30$, $SD = 1.30$, SE = .29) to final ($M = 2.05$, $SD = 1.19$, SE = .27), $t(19) = 3.14$, $p = .005$, Cohen's $d = 0.60$, but older children's initial ($M = 2.09$, $SD = 1.12$, SE = .23) and final ($M = 1.91$, $SD = 1.12$, SE = .23) attributions were not significantly different. As Figure 8 illustrates, older children's *initial* attribution of intentions to Robbie were significantly higher than younger children's *initial* attributions, $t(42) = 1.72$, $p = .03$, but older and younger children's *final* attributions were not significantly different. As noted earlier, comparing responses with Study 1, it appears that older children may have already been more keen on noticing that the robot was moving seemingly without any external control. For younger children, additional evidence from the "surveillance" video may have served to direct their attention to the same idea, that Robbie was moving by itself. In contrast, older children may have already established a level of belief about the robot's autonomy that was relatively less susceptible to revision.

IMPRESSIONS ABOUT ROBBIE

On average, children claimed to like Robbie "a lot" ($M = 2.84$, $SD = 0.43$), they tended to claim that Robbie was "not at all" or "not much" scary ($M = 0.45$, $SD = 0.85$), and they tended to think that Robbie was "a little bit" cute ($M = 1.95$, $SD = 0.94$). Liking was not significantly predicted by either age or sex. Sex (dummy coded: girls = 0, boys =

1) predicted judgments of scariness, ($\beta = -.70, p = .01$), such that girls tended to think Robbie was a bit scary more than boys, $R^2 = .15$ $F(1, 42) = 7.37, p = .01$. Age predicted attributions of cuteness ($\beta = -.31, p = .02$), such that, with increased aged, children were less likely to say that Robbie was cute, $R^2 = .12, F(1, 42) = 5.52, p = .02$.

Multiple regression analysis was used to determine whether each of these judgments was predicted by children's initial attributions of each of the 16 original psychological attributions. A forward-selection stepwise regression on children's ratings of whether they liked Robbie resulted in a model with children's attributions of whether Robbie could feel happy ($\beta = .19, p = .02$) as the only significant predictor, $R^2 = .12, F(1, 41) = 5.78, p = .02$. A forward-selection stepwise regression on children's ratings of whether they thought Robbie was scary did not reveal any significant predictors. A forward-selection stepwise regression on children's ratings of whether they thought Robbie was cute resulted in a model with children's attributions of whether Robbie could feel loved ($\beta = .50, p < .001$) as the only significant predictor, $R^2 = .26, F(1, 41) = 14.51, p < .001$.

In addition, a series of multiple regressions was used to determine whether children's psychological attributions varied as a function of whether children liked Robbie, thought Robbie was scary, or whether they thought Robbie was cute. A forward-selection stepwise regression on children's attributions of negative emotions (average of initial and final) resulted in a model with children's cuteness rating ($\beta = .29, p = .03$) as the only significant predictor, $R^2 = .11, F(1, 41) = 4.80, p = .03$. Similarly, a forward-selection stepwise regression on children's attributions of positive emotions (average of

initial and final) resulted in a model with children's cuteness rating ($\beta = .42, p = .001$) as the only significant predictor, $R^2 = .24, F(1, 40) = 12.83, p = .001$. Finally, a forward-selection stepwise regression on children's attributions of cognition (average of initial and final) resulted in a model with children's cuteness rating ($\beta = .33, p = .01$) as the only significant predictor, $R^2 = .15, F(1, 42) = 7.14, p = .01$. Thus, the cuter children thought Robbie was, the more they attributed negative and positive emotions and cognition to the robot. Children's judgments of intentions did not vary as a function of liking, scariness, or cuteness ratings.

PARENT QUESTIONNAIRE

Psychological Attributions

A series of multiple regressions was used to determine whether children's psychological attributions varied as a function of each of the five items from the parent questionnaire: whether children had seen a robot in person before (dummy coded: no = 0, yes = 1), children's experience with robot-like toys, frequency viewing media with robot characters, whether parents conversed with their children about how robots work, and experience with pets/animals. A forward-selection stepwise regression on children's attributions of negative emotions (average of initial and final) did not reveal any significant predictors. None of the items significantly predicted attributions of positive emotions (average of initial and final) either. However, a forward-selection stepwise regression on children's attributions of cognition (average of initial and final) resulted in a model with children's experience with robot-like toys ($\beta = -.25, p = .02$) as the only

significant predictor, $R^2 = .12, F(1, 42) = 5.65, p = .02$. The more experience children had with robotic toys resembling animals or people, the less they tended to attribute cognitive abilities to Robbie the Robot. Attribution of intentions was not predicted by any of the items from the parent questionnaire.

Moral Accountability

None of the five items from the parent questionnaire emerged as significant predictors of children's moral accountability judgments.

Impressions

A forward-selection stepwise regression on children's judgments of how much they liked Robbie did not reveal any significant predictors. A forward-selection stepwise regression on children's judgments of the robot's scariness resulted in a model with children's previous experience having seen a robot in person ($\beta = -.65, p = .01$) as the only significant predictor, $R^2 = .14, F(1, 42) = 6.86, p = .01$. Children who had previously seen a robot were less likely to think that Robbie was scary. A forward-selection stepwise regression on children's judgments of the robot's cuteness resulted in a model with children's previous experience with robotic toys ($\beta = -.27, p = .04$) as the only significant predictor, $R^2 = .10, F(1, 42) = 4.66, p = .04$. Thus, with more experience playing with robotic toys, children were less likely to rate Robbie as cute.

SUMMARY

As mentioned earlier, Study 2 was designed to test the prediction that making the robot appear to be a moral agent would specifically increase children's attributions of psychological agency but not experience to the robot. The results offer little support for this hypothesis, as children's final psychological attributions remained relatively unchanged from their initial ones, after seeing Robbie the Robot destroy the tower. The only exception was that 5-year-olds' judgments for the intentions item (whether the robot could do things on purpose) increased from initial to final, replicating Study 1 results, but this effect was relatively small. In contrast, 7-year-olds' judgments of intentions remained unaltered, perhaps partly because their initial responses were already higher than they were in Study 1.

It was surprising that the intentions item had a unique factor loading separate from the other seven items that were also designed to tap into children's ideas about agency. In addition, the intentions item consistently emerged as a significant predictor of moral accountability judgments, as it did in Study 1. It appears that the wording "do things on purpose" uniquely tapped into some concept of intentionality, perhaps even moral intentionality, that none of the other items seemed to capture. In future research, rather than rely solely on explicit answers to questions about psychological properties that may be sensitive to linguistic interpretation, it might fruitful to introduce behavioral measures of anthropomorphism to gain a better understanding of children's implicit and explicit beliefs about the robot's psychological status (Woolley, 2006; Severson & Carlson, 2010). In addition, behavioral measures of trust might…

One noteworthy finding from Study 2 was that children's judgments of whether Robbie the Robot was cute were associated with higher attributions of positive and negative emotions and cognition to Robbie. This result is consistent with claims made by Sherman and Haidt (2011) that perceptions of cuteness serve an evolved function to motivate social engagement, affecting the degree to which an entity is mentalized (i.e., how much they are imbued with mental states). In light of this, the fact that cuteness ratings in the current study were *not* associated with attributions of intentions to Robbie serves to highlight the fact that "doing things on purpose" was treated as a qualitatively different kind of psychological property than experiencing emotions or cognition. Indeed, prototypically cute entities, babies, are typically perceived as innocent and not able to "do things on purpose."

General Discussion

Smart technologies are already abundant in our lives. Soon, intelligent humanoid robots will be commonplace as well. These robots will take on roles in human lives as caretakers, soldiers, companions, and in countless other ways. As that future approaches, age-old questions from science fiction will finally be tested in reality as people begin to contemplate whether their humanlike robot peers are actually experiencing mental and emotional lives, and whether they should be regarded with the same moral consideration granted to other humans and other animals; the difference being that these robots will have a clear status as technological artifacts and not living creatures. Kahn, and colleagues (2011) have suggested that with the rise of personified technologies, a new ontological category is emerging in people's classification of the world; a unique category of entities that are considered non-biological, yet psychological and worthy of moral consideration. In particular, children developing in this new technological landscape will likely have a unique stance about the status of such personified technologies, especially regarding humanoid robot companions. If we are to create a harmonious coexistence of robots with humans, it is important to begin to understand how young children will come to view such robots. This dissertation attempts to shed some light on young children's emerging views about the psychological and moral status of robots.

Aside from replicating previous research showing that children's emerging views about robots consist of a complex set of attributions that cross traditional boundaries of prototypical living and non-living entities (Kahn, 2011), the current studies are the first to

demonstrate a clear link between children's attributions of psychological agency, in particular the capacity to exhibit intentional behavior, and moral accountability to a humanoid robot. Children's judgments about whether Robbie the Robot destroyed the tower intentionally ("on purpose") predicted their judgments of whether Robbie deserved to get in trouble (Studies 1 and 2) and whether Robbie was judged as having been bad (Study 2). A similar relationship between intentionality and moral accountability has been demonstrated through previous studies where the agent is a human (e.g., Cushman, Sheketoff, Wharton, & Carey, 2013), but it is important to note that because the agents presented to children in such studies are people, they are presumed by default to have the mental faculties necessary for intentional action. In contrast, in the current set of studies, the actor in question, a robot, was not automatically granted the full set of mental abilities a child might automatically grant to a person. Thus, these studies also provided the unique opportunity to assess a slightly different problem: whether beliefs about the existence of mental states in an entity predict children's judgments of moral accountability. Indeed, judgments of Robbie the Robot's moral accountability (including judgments of blameworthiness, deservedness of punishment, and naughtiness) were consistently predicted not just by children's judgments of whether the robot intended to destroy the tower but also by children's stance about whether the robot even had the ability for intentional behavior.

In addition, findings from Study 1 join a new line of research demonstrating that increasing the apparent autonomy of a robot (namely by increasing cues to self-initiated action) affects whether the robot is perceived as having a mental life and whether it is

considered worthy of moral concern (Gary, 2014; Gary & Chernyak 2013; Somanader, Saylor, & Levin, 2011). Findings from the current study expand upon this work by providing evidence that a robot's apparent autonomy can affect whether a young child sees the robot as morally responsible for causing harm. Specifically, Study 1 demonstrates that cues to autonomous motion increase young children's judgments of whether a humanoid robot is seen as morally accountable (blameworthy, naughty, and deserving of punishment), and this relationship is mediated specifically by children's attributions of intentionality to the robot.

Related to this finding, recent dissertation work by Gary (2014) demonstrates that the framing of humanoid robot's source of agency (i.e., whether is appears autonomous or controlled by a person) also affects whether adults credit the robot for making positive contributions to a task, and that this relationship is mediated by attributions of psychological agency to the robot. Taken together, findings from Gary's (2014) work and the current study demonstrate that cues to self-initiated action affect adults' and young children's attributions of praise and blame via the attribution of psychological agency. In other words, when a robot's source of control appears to be internal, this creates the illusion that the robot has some intentionality, which then elicits the idea that the robot should be treated as an individual worthy of praise if it causes good outcomes and deserving of punishment if it causes harm.

As robots become increasingly sophisticated and technologically autonomous, it is possible that perceptions of the source of their control will continue to shift away from the external, and more toward the robot as an individual being. Ultimately, this may

result in perceptions of some robots as essentially human in their psychological agency. Of course, this will also depend on individual differences in beliefs about what constitutes being human. For example, Khan et al. (2007) suggest that the degree to which a person grants autonomy to robots may ultimately depend partly on whether the person grants autonomy to themselves or to other people. As Khan et al. (2007) point out, historically there has been philosophical debate about whether we, as humans, have full autonomy or whether our actions are mechanistically determined by forces beyond our control. For individuals, beliefs about the source and scope of human agency may serve to define parameters for what we think would be ultimately possible for a robot's agency and control.

Interestingly, in this dissertation, some of the children's responses reflected the idea that a person and/or a computer is usually in control of the robot's actions, but that at times, the robot is capable of acting independently of this control. For example, one 7-year-old boy stated that Robbie is "controlled by a computer, which is controlled by a person, so it's basically controlled by a person. But if he ever has free time it's controlled by him." Another 7-year-old boy explained that "sometimes the electricity in his head fights with the cameras so the cameras fight against the electricity making the cameras go fuzzy so then he does the thing that he wants to do." Responses such as these suggest that, even with some understanding that robots can be controlled remotely by a person or by programming (Scaife & van Duuren, 1995), children may still be inclined to grant some mental capacity, perhaps even free will, to robots. In future studies, it would be interesting to measure such individual differences, for example beliefs about free will

(Kushnir, Gopnik, Chernyak, Seiver, & Wellman, 2015), souls (Richert & Harris, 2008), and intuitive dualism (Bloom, 2004), and to assess if these beliefs predict a child's willingness to attribute psychological agency to a robot.

Returning to issue of autonomy and moral accountability, it may be important to consider that as technological agents become more capable of causing harm to humans (accidental or not), people will likely start to attribute more mental life to these machines (Kahn et al., 2012). Related to this, currently there are heated debates about the consequences of introducing autonomous vehicles onto the roads and about whether autonomous drones can or should be allowed to make moral judgments about alternative courses of action, such as whether or not to fire a missile into a populated region. Policy about such issues will likely be influenced by people's general perceptions about the psychological and moral status of such intelligent agents (Kahn et al., 2012).

Regarding autonomous vehicles, an interesting issue is how people will deal with accidents that will inevitably occur (even though these accidents will likely be more infrequent than when humans are behind the wheel). As demonstrated in moral typecasting work (Gray & Wegner, 2009), people tend to search for responsible agents when someone is harmed, especially if the victim is perceived to be innocent. This poses an interesting quandary with respect to autonomous vehicles, because if an accident were to occur, who would be to blame? Certainly, this question is already posing unique challenges to policy-makers and insurance companies regarding legal action and monetary compensation for such incidents. But, perhaps just as important is that, given the human proclivity to seek justice by assigning blame to an agent, people are likely to

begin treating autonomous technologies as morally accountable for their actions, even if that blame is simultaneously assigned to a human creator or programmer as well.

Indeed, findings from Study 1 show that children are not only inclined to blame a robot for causing harm, but they are just as willing to blame the robot along with a person as "coconspirators" when children understand that the person is controlling the robot's actions. Interestingly, in Study 2, with no one else even in the room with the robot, one 7-year-old girl's response to the experimenter's question of whether it was Robbie's fault that the tower was destroyed was: "That's a hard one but it's because someone back there was messing with the controls of him because I think he's being controlled… I'd say about fifty percent [Robbie's fault] and the other person has fifty percent."

As autonomous technological agents start making their way into day-to-day interactions with people, it will be important to consider methods of reducing blame and mistrust through psychologically informed design of these agents. A recent experiment by Waytz and Epley (2014) using a driving simulator suggests that in the case of autonomous vehicles, incorporating anthropomorphic features such as giving the car a gendered voice and a name may actually serve to increase trust assigned to the car. Thus, it may be that as artificial agents become more autonomous and seemingly capable of intentional behaviors and psychological agency, one way to reduce the inclination to blame them will be to introduce features that add other dimensions of personality and increase attributions of emotional experience as well.

In order to bolster trust, companionship, and a harmonious coexistence with emerging technological agents, it may be most important to consider factors that promote empathy.

As robots become increasingly autonomous and capable of committing meaningful and consequential acts in the physical and social world, their actions will carry moral implications that, in the absence of a human being to blame, will be turned on them. Future research should explore under what conditions people are likely to empathize with and forgive an autonomous robot after causing harm. More generally, gaining a more complete understanding of the causes and consequences of different forms of anthropomorphism (accounting various kinds of mental states that might be attributed) will help guide robot designers to create life-like entities that people will be able to connect with at a meaningful level. In addition, as we enter a world where humans and robots increasingly start to coexist, continued research at this historically unique frontier may ultimately help us gain a better understanding of what it means to be human.

Appendix A: Psychological Attribution Questions Used in Study 1

1. Thoughts: Can [entity] think?
2. Intentions: Can [entity] do things on purpose?
3. Emotions: Can [entity] feel things like happy or sad?
4. Sensations: If someone poked [entity], would [entity] feel it?

Appendix B: Psychological Attribution Questions Used in Study 2

1. Can Robbie control his own actions?
2. Can Robbie make plans?
3. Can Robbie make decisions?
4. Can Robbie do things on purpose?
5. Can Robbie remember things?
6. Can Robbie think?
7. Can Robbie have ideas?
8. Does Robbie have a mind?
9. Can Robbie feel upset?
10. Can Robbie feel scared?
11. Can Robbie feel hurt?
12. Can Robbie feel sad?
13. Can Robbie feel loved?
14. Can Robbie feel excited?
15. Can Robbie feel good?
16. Can Robbie feel happy?

Table 1: Study 2 Factor Loadings for Initial Psychological Attributions

Note: Coefficients > .40 are in bold.

	Pattern Coefficients				Structure Coefficients			
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 1	Factor 2	Factor 3	Factor 4
Rob1_Upset	.087	**.756**	.000	.240	.298	**.798**	-.242	.292
Rob1_Scared	.243	**.756**	.168	-.319	.394	**.760**	-.125	-.277
Rob1_Hurt	-.149	**.810**	-.102	.100	.127	**.801**	-.248	.167
Rob1_Sad	-.111	**.800**	-.095	.107	.159	**.800**	-.256	.172
Rob1_Loved	**.408**	.251	**-.449**	-.180	**.685**	**.471**	**-.689**	-.157
Rob1_Excited	.317	.211	**-.598**	-.166	**.649**	**.445**	**-.789**	-.138
Rob1_Good	.314	.219	**-.625**	-.140	**.659**	**.460**	**-.816**	-.111
Rob1_Happy	-.078	.196	**-.850**	-.137	.362	.384	**-.861**	-.091
Rob1_Control	-.018	-.218	**-.890**	.068	.315	.012	**-.828**	.084
Rob1_Plans	**.714**	.064	-.085	.167	**.766**	.303	**-.426**	.155
Rob1_Decisions	.381	.088	**-.489**	.358	**.615**	.350	**-.694**	.372
Rob1_Purpose	.094	.221	.146	**.826**	.071	.270	.017	**.835**
Rob1_Remember	**.815**	.012	.087	-.281	**.787**	.204	-.270	-.305
Rob1_Think	**.847**	.046	-.007	.086	**.861**	.298	**-.400**	.067
Rob1_Ideas	**.457**	.359	-.275	.006	**.683**	**.562**	**-.572**	.030
Rob1_Mind	**.825**	-.228	-.114	.125	**.807**	.048	**-.427**	.091

Figure 1. The humanoid robot, Nao (Adebaran Robotics). Picture of “Robbie the Robot” presented to children during the property attribution interview in Studies 1 and 2.

Figure 2. Pictures of the items used for the property attribution interview in Study 1.

 Car	 Computer
 Tree	 Baby

Figure 3. Four-point scale children used to rate responses to attribution questions and moral accountability questions in Studies 1 and 2.

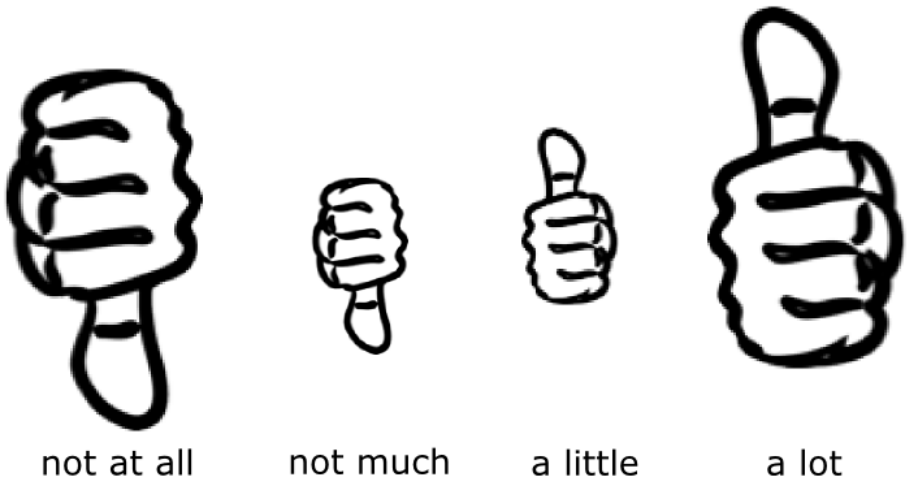


Figure 4. Study 1 average psychological attribution to Robbie the Robot, a computer, a tree, a car, and a baby. Error bars denote standard error of the mean.

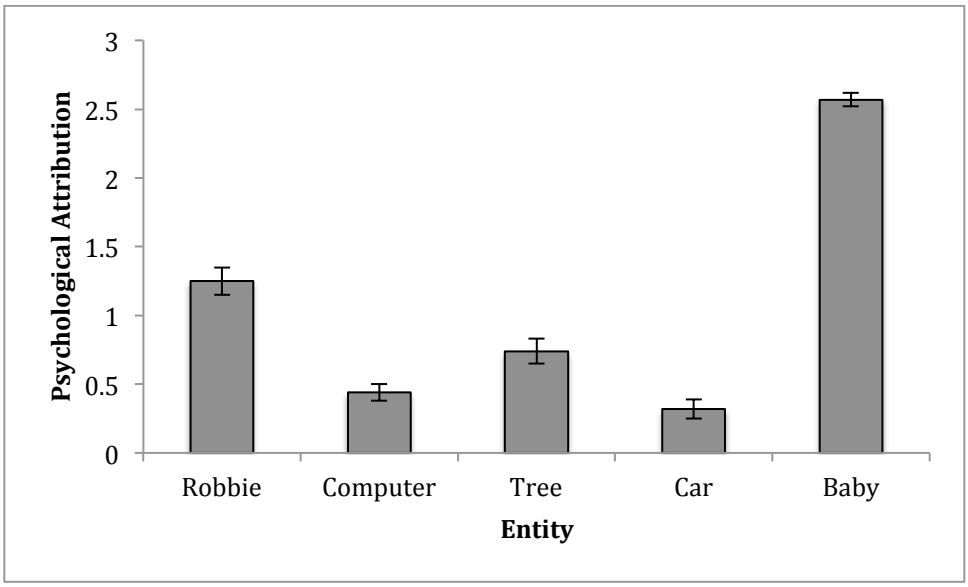


Figure 5. Study 1 moral accountability judgments as a function of condition (autonomous vs. controlled) and agent in question (Robbie the Robot vs. Experimenter 2). Error bars denote standard error of the mean.

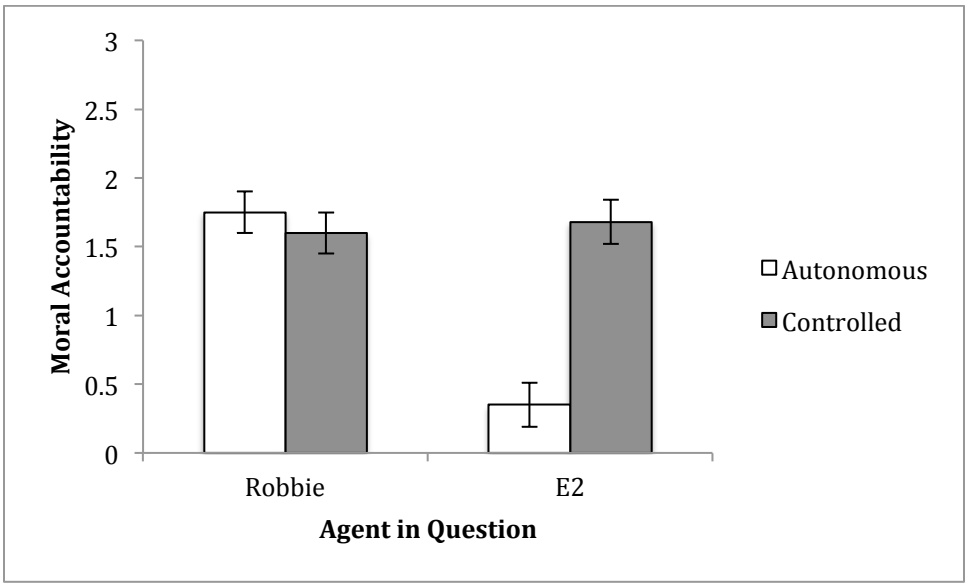
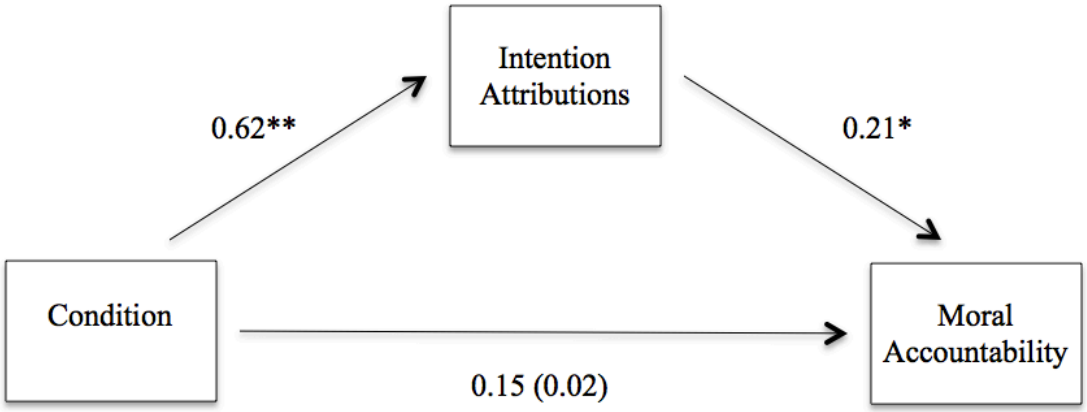


Figure 6. The relationship between condition (autonomous vs. controlled) on judgments of Robbie the Robot's moral accountability as mediated by attributions of intentions to the robot (Study 1).
*$p < .05$, **$p < .01$.



Estimate of Indirect Effect: β = 0.13 (95% CI: 0.003, 0.381)

Reduction in Model: 86.67%

Figure 7. Study 1 average psychological attribution to Robbie the Robot as a function of attribution type (Thoughts = "Can Robbie Think?"; Intentions = "Can Robbie do things on purpose?"; Emotions = "Can Robbie feel things like happy or sad?"; Sensations = "If someone poked Robbie, would Robbie feel it?"), condition (Autonomous = no apparent external control, Controlled = actions explicitly controlled by second experimenter), and time (initial = before tower destruction; final = after tower destruction). Error bars denote standard error of the mean.

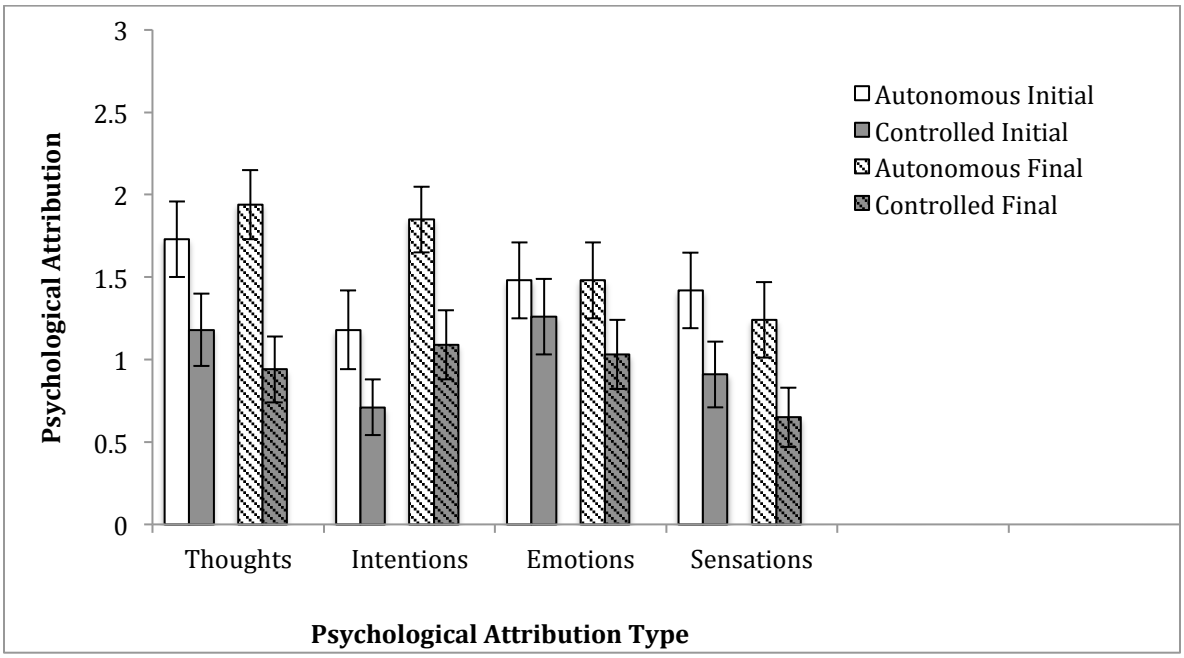
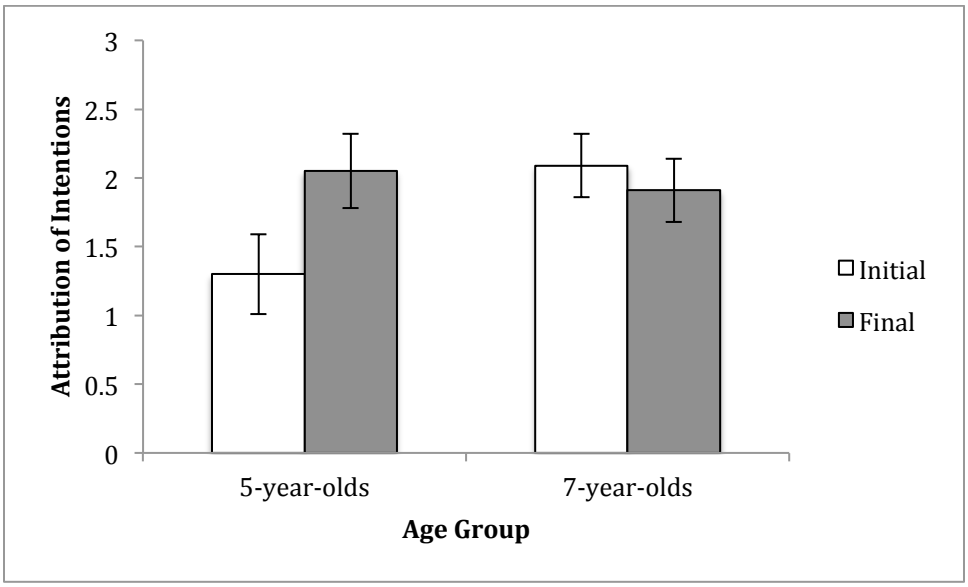


Figure 8. Study 2 attributions of intentions ("Can Robbie do things on purpose?") by age (5-year-olds vs. 7-year-olds) and time (initial = before tower destruction; final = after tower destruction). Error bars denote standard error of the mean.



References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*, 1173.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York, NY: Oxford University Press.

Bentham, J., & Browning, J. (1843). *The works of Jeremy Bentham*. London: Simpkin, Marshall, & Co.

Bertenthal, B. I. (1993). Infants' perception of biomechanical motions: Intrinsic image and knowledge-based constraints. In C. Granrud (Ed.), *Visual perception and cognition in infancy* (pp. 175-214). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. New York, NY: Basic Books.

Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. *Procedings of the IROS '99: IEEE/RSJ International Conference on Intelligent Robots and Systems*, *2*, 858-863.

Burghardt, Gordon M (1985). Animal awareness: Current perceptions and historical perspective. *American Psychologist, 40*, 905–919.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125*, 1839-1849.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. New York, NY: Oxford University Press.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*, 6-21.

Dennett, D. C. (2008). *Kinds of minds: Towards an understanding of consciousness*. New York, NY: Basic Books.

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating Social Connection Through Inferential Reproduction Loneliness and Perceived Agency in Gadgets, Gods, and Greyhounds. *Psychological Science*, *19*, 114-120.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*, 143-155.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*, 864-886.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*, 493-501.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, *8*, 396-403.

Gary, H. E., Chernyak, N. (2013). *Self-generated, goal-directed movement predicts children's moral regard and prosocial behavior*. Poster presented at the meeting of the Cognitive Development Society, Memphis.

Gary, H. E. (2014). *Adults' Attributions of Psychological Agency, Credit, and Fairness to a Humanoid Social Robot* (Doctoral dissertation, University of Washington).

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *Neuroimage*, *35*, 1674-1684.

Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*, 165-193.

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

Gray, K., Jenkins, A. C., Heberlein, A. S., & Wegner, D. M. (2011). Distortions of mind perception in psychopathology. *Proceedings of the National Academy of Sciences*, *108*, 477-479.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*, 505.

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

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*, 101-124.

Guthrie, S. (1993). *Faces in the clouds: A new theory of religion*. New York, NY: Oxford University Press.

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*, 451-470.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26(1), 30-39.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation in preverbal infants. *Nature*, *450*, 557-560.

Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language*, *7*, 120–144.

Hood, B. 2008. *Supersense: Why We Believe in the Unbelievable*. New York, NY: HarperCollins Publishers.

Hume, D. (1957). *The natural history of religion*. Stanford, CA: Stanford University Press. (Original work published 1757).

Hutson, M. (2012). *The 7 laws of magical thinking: How irrational beliefs keep us happy, healthy, and sane*. New York, NY: Hudson Street Press.

Inagaki, K., & Hatano, G. (1987). Young children's spontaneous personification as analogy. *Child Development*, *58*, 1013-1020.

Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child Development*, *78*, 1675-1688.

Johnson, M. H., & Morton, J. (1991). *Biology and cognitive development: The case of face recognition*. Oxford, UK: Basil Blackwell.

Kahn, P. H., Gary, H. E., & Shen, S. (2013). Children's social relationships with current and near-future robots. *Child Development Perspectives*, *7*(1), 32-37.

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., Ruckert, J. H., & Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, *48*, 303-314.

Kahn, P.H., Jr., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., Freier, N., & Severson, R.L. (2012). *Do people hold a humanoid robot morally accountable for the harm it causes?* Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction, 33-40.

Kahn Jr, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., ... & Gill, B. (2011, March). The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction* (pp. 159-160). ACM.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*, 1074.

Killen, M., & Smetana, J.G. (in press). Morality: Origins and development. In M. Lamb & C. Garcia-Coll (Eds.), *Handbook of child psychology and developmental science, Vol. 3, 7th edition*, Editor-in-Chief, R. M. Lerner. NY: Wiley-Blackwell.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190-194.

Kushnir, T., Gopnik, A., Chernyak, N., Seiver, E., & Wellman, H. M. (2015). Developing intuitions about free will between ages four and six. *Cognition*, *138*, 79-101.

Kwan, V. S., Gosling, S. D., & John, O. P. (2008). Anthropomorphism as a special case of social perception: A cross-species social relations model analysis of humans and dogs. *Social Cognition*, *26*, 129-142.

Lagattuta, K. H., Sayfan, L., & Monsour, M. (2011). A new measure for assessing executive function across a wide age range: children and adults find happy-sad more difficult than day-night. *Developmental Science*, *14*, 481-489.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect theory of mind and moral judgment. *Psychological Science*, *17*(5), 421-427.

Low, P., Panksepp, J., Reiss, D., Edelman, D., Van Swinderen, B., Low, P., & Koch, C. (2012, July 7). *The Cambridge declaration on consciousness*. Retrieved from *http://fcmconference. org/img/CambridgeDeclarationOnConsciousness.pdf*

Mataric, M. J. (2000). Getting humanoids to move and imitate. *IEEE Intelligent Systems*, *15*(4), 18-24.

Melson, G. F., Kahn Jr, P. H., Beck, A., & Friedman, B. (2009). Robotic pets in human lives: Implications for the human–animal bond and for human relationships with personified technologies. *Journal of Social Issues*, *65*, 545-567.

Mithen, S. J. (1996). *The prehistory of the mind: A search for the origins of art, religion, and science*. London: Thames and Hudson.

Newman, G. E., Keil, F. C., Kuhlmeier, V. A., & Wynn, K. (2010). Early understandings of the link between agents and order. *Proceedings of the National Academy of Sciences*, *107*, 17140-17145.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: The MIT Press.

Piaget, J. (1929). *The child's conception of the world*. London: Kegan Paul.

Richert, R. A., & Harris, P. L. (2008). Dualism revisited: Body vs. mind vs. soul. *Journal of Cognition and Culture*, *8*, 99-115.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169-192.

Robins, B., Dautenhahn, K., Te Boekhorst, R., & Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?. *Universal Access in the Information Society, 4*, 105-120.

Roy, N., Baltus, G., Fox, D., Gemperle, F., Goetz, J., Hirsch, T., Margaritis, D., Montemerlo, M., Pineau, J., Schulte, J., & Thrun, S. (2000, July). Towards personal service robots for the elderly. In *Workshop on Interactive Robots and Entertainment*, *25*, 184.

Scaife, M., & van Duuren, M. (1995). Do computers have brains? What children believe about intelligent artifacts. *British Journal of Developmental Psychology*, *13*, 367-377.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1), 13-24.

Scassellati, B. (2007). How social robots will help us to diagnose, treat, and understand autism. In *Robotics Research,* 552-563.

Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, *23*, 1099-1103.

Sharkey A, & Sharkey N. (2011). Children, the elderly, and interactive robots: Anthropomorphism and deception in robot care and companionship. *IEEE Robotics and Automation Magazine*, 18, 32–38.

Sherman, G. D., & Haidt, J. (2011). Cuteness and disgust: the humanizing and dehumanizing effects of emotion. *Emotion Review*, *3*, 245-251.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010, March). No fair!! an interaction with a cheating robot. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on* (pp. 219-226). IEEE.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological methods*, *7*, 422.

Somanader, M. C., Saylor, M. M., & Levin, D. T. (2011). Remote control and children's understanding of robots. *Journal of Experimental Child Psychology*, *109*, 239-247.

Subbotsky, E. V. (1993). *Foundations of the mind: Children's understanding of reality*. Cambridge, MA: Harvard University Press.

Wada, K., & Shibata, T. (2007). Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house. *Robotics, IEEE Transactions*, *23*, 972-980.

Walter, W. G. (1950). An imitation of life. *Scientific American*, *182*(5), 42-45.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*, 1301-1303.

Waytz, A., Cacioppo, J.T., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*, 219-232.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*, 410.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, *66*, 297–331.

Willard, A. K., & Norenzayan, A. (2013). Cognitive biases explain religious belief, paranormal belief, and belief in life's purpose. *Cognition*, *129*, 379-391.

Wimmer, H., Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.

Woolley, J. D. (1997). Thinking about fantasy: Are children fundamentally different thinkers and believers from adults?. *Child Development*, *68*, 991-1011.

Woolley, J. D. (2006). Verbal–behavioral dissociations in development. *Child Development*, *77*, 1539-1553.

Zaitchik, D., & Solomon, G. E. (2008). Animist thinking in the elderly and in patients with Alzheimer's disease. *Cognitive Neuropsychology*, *25*(1), 27-37.

Zawieska, K., Duffy, B. R., & Sprońska, A. (2012). Understanding anthropomorphisation in social robotics. *Pomiary Automatyka Robotyka*, *11*, 78-82.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development*, *67*, 2478-2492.