What is a human?

Toward psychological benchmarks in the field of human–robot interaction

Peter H. Kahn, Jr., Hiroshi Ishiguro, Batya Friedman, Takayuki Kanda, Nathan G. Freier, Rachel L. Severson and Jessica Miller
University of Washington / Osaka University and Advanced Telecommunications Research / University of Washington / Advanced Telecommunications Research / University of Washington

In this paper, we move toward offering psychological benchmarks to measure success in building increasingly humanlike robots. By psychological benchmarks we mean categories of interaction that capture conceptually fundamental aspects of human life, specified abstractly enough to resist their identity as a mere psychological instrument, but capable of being translated into testable empirical propositions. Nine possible benchmarks are considered: autonomy, imitation, intrinsic moral value, moral accountability, privacy, reciprocity, conventionality, creativity, and authenticity of relation. Finally, we discuss how getting the right group of benchmarks in human–robot interaction will, in future years, help inform on the foundational question of what constitutes essential features of being human.

Keywords: authenticity of relation, autonomy, creativity, human–robot interaction, imitation, morality, privacy, psychological benchmarks, reciprocity, robot ethics

In computer science, benchmarks are often employed to measure the relative success of new work. For example, to test the performance of a new database system one can download a relevant benchmark (e.g., from www.tpc.org): a dataset and a set of queries to run on the database. Then one can compare the performance of the system to other systems in the wider community. But in the field of human–robot interaction, if one of the goals is to build increasingly humanlike robots, how do we measure success? In this paper, we focus on the psychological aspects of this question. We first set the context in terms of humanoid robots, and then

Interaction Studies 8:3 (2007), **363–390**.
ISSN 1572–0373 / E-ISSN 1572–0381

distinguish between ontological and psychological claims about such humanoids. Then we offer nine possible psychological benchmarks for consideration. Finally, we discuss how getting the right group of benchmarks in human–robot interaction will, in future years, help inform on the foundational question of what constitutes essential features of being human.

Why build humanlike robots?

We would like to acknowledge that there are some good reasons *not* to have the goal to build humanlike robots. One reason, of course, is that in many forms of human–robot interaction there is nothing gained functionally by using a humanoid (e.g., assembly line robots). There are also contexts where the humanlike form may work against optimal human–robot interaction. For example, an elderly person may not want to be seen by a robot with a humanlike face when being helped to the bathroom. In addition, humans may dislike a robot that looks human but lacks a human behavioral repertoire, part of a phenomenon known as the *uncanny valley* (Dautenhahn, 2003; MacDorman, 2005).

That said, there are equally good reasons to aim to build humanlike robots. Functionally, for example, human–robot communication will presumably be optimized in many contexts if the robot conforms to humanlike appearance and behavior, rather than asking humans to conform to a computational system (Ishiguro, 2004; Kanda, Hirano, Eaton, & Ishiguro, 2004; Kanda, Ishiguro, Imai, & Ono, 2004; Minato, Shimada, Ishiguro, & Itakura, 2004). Psychological benefits could also accrue if humans 'kept company' with robotic others (Kahn, Freier, Friedman, Severson, & Feldman, 2004). And perhaps no less compelling, benefits or not, there is the long-standing human desire to create artifactual life, as in stories of the Golem from the 16th century.

Distinguishing ontological and psychological claims

Two different types of claims can be made about humanoid robots at the point when they become (assuming it possible) virtually human-like. One type of claim, ontological, focuses on what the humanoid robot actually is. Drawing on Searle's (1990) terminology of "Strong and Weak AI," the strong ontological claim is that at this potentially future point in technological sophistication, the humanoid actually becomes human. The weak ontological claim is that the humanoid only appears to become human, but remains fully artifactual (e.g., with syntax but not semantics). A second type of claim, the psychological, focuses on what people attribute to the

fully humanlike robot. The strong psychological claim is that people would conceive of the robot as human. The weak psychological claim is that people would conceive of the robot as a machine, or at least not as a human.

In turn, there are four possible combinations of the ontological and psychological claims. *Case 1.* The robot (ontologically speaking) becomes a human, and people (psychologically speaking) believe the robot is a human, and act accordingly. *Case 2.* The robot (ontologically speaking) becomes a human, but people (psychologically speaking) neither believe a robot can become human nor act accordingly. *Case 3.* The robot cannot (ontologically speaking) become a human, but people (psychologically speaking) believe the robot is a human, and act accordingly. And *Case 4.* The robot cannot (ontologically speaking) become a human, and people (psychologically speaking) neither believe a robot can become human nor act accordingly. In Cases 1 and 4, people's psychological beliefs and actions would be in accord with the correct ontological status of the robot, but in Cases 2 and 3 they would not.

Thus, there is an important distinction between claims regarding the ontological status of humanoid robots and the psychological stance people take toward them. Much debate in cognitive science and artificial intelligence has centered on ontological questions: Are computers as we can conceive of them today in material and structure capable of becoming conscious? (Hofstadter & Dennett, 1981). And regardless of where one stands on this issue — whether one thinks that sometime in the future it is possible to create a technological robot that actually becomes human, or not — the psychological question remains. Indeed, in terms of societal functioning and wellbeing, the psychological question is at least as important as the ontological question.

Toward psychological benchmarks

The issue at hand then becomes, psychologically speaking, how do we measure success in building humanlike robots? One approach might be to take findings from the psychological scientific disciplines, and seek to replicate them in human–robot interaction. The problem here is that there must be at least tens of thousands of psychological findings in the published literature over the last 50 years. In terms of resources, it is just not possible to replicate all of them. Granted, one could take a few hundred or even a few thousand of some of the findings, and replicate them on human–robot interaction. But, aside from good intuitions and luck, on what bases does one choose which studies to replicate? Indeed, given that human–robot interaction may open up *new* forms of interaction, then even here the existing corpus of psychological research comes up short. Thus in our view the field of HRI would be well-served by establishing psychological benchmarks.

Our first approximation for what we mean by psychological benchmarks is as follows: categories of interaction that capture conceptually fundamental aspects of human life, specified abstractly enough so as to resist their identity as a mere psychological instrument (e.g., as in a measurement scale), but capable of being translated into testable empirical propositions. Although there has been important work on examining people's humanlike responses to robots (e.g., Dautenhahn, 2003; Aylett, 2002; Bartneck, Nomura, Kanda, Suzuki, & Kato, 2005; Breazeal, 2002; Kaplan, 2001; Kiesler & Goetz, 2002) and on common metrics for task-oriented human–robot interaction (Steinfeld, Fong, Kaber, Lewis, Scholtz, Schultz, & Goodrich, 2006), we know of no literature in the field that has taken such a direct approach toward establishing psychological benchmarks.

Nine psychological benchmarks

With the above working definition in hand, we offer the following nine psychological benchmarks. Some of the benchmarks are characterized with greater specificity than others, and some have clearer measurable outcomes than others, given the relative progress we have made to date. We also want to emphasize that these benchmarks offer only a partial list of possible contenders; and indeed some of them may ultimately need to be cast aside, or at least reframed. But as a group they do help to flesh out more of what we mean by psychological benchmarks, and why they may be useful in future assessments of human–robot interaction.

1. Autonomy

A debated issue in the social sciences is whether humans themselves are autonomous. Psychological behaviorists (Skinner, 1974), for example, have argued that people do not freely choose their actions, but are conditioned through external contingencies of reinforcement. Endogenous theorists, as well, have contested the term. For example, sociobiologists have argued that human behavior is genetically determined, and that nothing like autonomy need be postulated. Dawkins (1976) writes, for example: "We are survival machines — robot vehicles blindly programmed to preserve the selfish molecules known as genes" (p. ix).

In stark contrast, moral developmental researchers have long proposed that autonomy is one of the hallmarks of when a human being becomes moral. For example, in his early work, Piaget (1932/1969) distinguished between two forms of social relationships: heteronomous and autonomous. Heteronomous relationships are constrained by a unilateral respect for authority, rules, laws, and the social order; in contrast, autonomous relationships — emerging, according to Piaget in

middle childhood — move beyond such constraints and become (largely through peer interaction) a relationship based on equality and mutual respect. Along similar lines, Kohlberg and his colleagues (Kohlberg, 1984) proposed that only by the latter stages of moral development (occurring in adolescence, if ever) does moral thinking differentiate from fear of punishment and personal interest (stages 1 and 2) as well as conventional expectations and obedience to social systems (stages 3 and 4) to become autonomous (stages 5 and 6).

Autonomy means in part independence from others. For it is only through being an independent thinker and actor that a person can refrain from being unduly influenced by others (e.g., by Neo-Nazis, youth gangs, political movements, and advertising). But as argued by Kahn (1999) and others, autonomy is not meant as a divisive individualism, but is highly social, developed through reciprocal interactions on a microgenetic level, and evidenced structurally in incorporating and coordinating considerations of self, others, and society. In other words, the social bounds the individual, and vice-versa.

Clearly the behavior of humanoid robots can and will be programmed with increasing degrees of sophistication to mimic autonomous behavior. But will people come to think of such humanoids as autonomous? Imagine, for example, the following scenario (cf. Apple's *Knowledge Navigator* video from 1987). You have a personal robot assistant at home that speaks through its interface with a voice that sounds about your age, but of the opposite gender. You come home from work and he/she (the robot) says: "Hey there, good to have you home, how did your meeting with Fred go today?" Assuming you have a history of such conversations with your robot, do you respond in a "normal" human way? Regardless, might he/she somehow begin to encroach on the relationship you have with your spouse? How about if he/she says, "Through my wireless connection, I read your email and you have one from your mom, and she really wants you to call her, and I think that should be your first priority this evening." Do you tell the robot: "It's not your role to tell me what to do." What if the robot responds, "Well, I was programmed to be an autonomous robot, that's what you bought, and that's what I am, and I'm saying your mom should be a priority in your life." What happens next? Do you grant the robot its claim to autonomy?

Such questions can help get traction on the benchmark. And answers will, in part, depend on clear assessments of whether, and if so how and to what degree, people attribute autonomy to themselves and other people.

2. Imitation

Neonates engage in rudimentary imitation, such as imitating facial gestures. Then, through development, they imitate increasingly abstract and complex phenomenon in ever broader contexts. Disputed in the field, however, is the extent to which

imitation can be categorized as being a highly active, constructive process as opposed to being rote (Gopnik & Meltzoff, 1998; Meltzoff 1995).

A partial account of the constructive process can be drawn from the work of James Mark Baldwin. According to Baldwin (1897/1973), there are three circular processes in a child's developing sense of self: the projective, subjective, and ejective. In the initial projective process a child does not distinguish self from other, but blindly copies others, without understanding. In the complementary subjective process, the child makes the projective knowledge his own by interpreting the projective imitative copy within, where "into his interpretation go all the wealth of his earlier informations, his habits, and his anticipations" (p. 120). From this basis, the child in the third process then ejects his subjective knowledge onto others, and "reads back imitatively into them the things he knows about himself" (p. 418). In other words, through the projective process the child in effect says, "What others are, I must be." Through the ejective process, the child in effect says, "What I am, others must be." Between both, the subjective serves a transformative function in what Baldwin calls generally the dialectic of personal growth. The important point here is that while imitation plays a central role in Baldwin's theory, it is not passive. Rather, a child's knowledge "at each new plane is also a real invention… He makes it; he gets it for himself by his own action; he achieves, invents it" (p. 106).

Given current trends in HRI research, it seems likely that humanoid robots will be increasingly designed to imitate people, not only by using language-based interfaces, but through the design of physical appearance and the expression of an increasing range of human-like behaviors (Akiwa, Sugi, Ogata, & Sugano, 2004; Alissandrakis, Nehaniv, Dautenhahn, & Saunders, 2006; Breazeal & Scassellati, 2002; Buchsbaum, Blumberg, Breazeal, & Meltzoff, 2005; Dautenhahn & Nehaniv, 2002; Yamamoto, Matsuhira, Ueda, & Kidode, 2004). One reason for designing robots to imitate people builds on the proposition that robotic systems can learn relevant knowledge by observing a human model. The implementation is often inspired by biological models (Dautenhahn & Nehaniv, 2002), including developmental models of infant learning (Breazeal & Scassellati, 2002; Breazeal, Buchsbaum, Gray, Gatenby, & Blumberg, 2005). Another reason for designing robots to imitate people is to encourage social interaction between people and robots (e.g., Yamamoto et al., 2004; Akiwa et al., 2004).

Thus one benchmark for imitation focuses on how successfully robots imitate people. Our point here is that, as with the benchmark of autonomy, it will be useful to have clear assessments of whether people believe that the robot imitates in a passive or active manner, and to compare those beliefs to whether people believe that humans imitate in an active or passive manner.

A second benchmark is perhaps even more interesting, and can be motivated by a fictional episode from the television program *Star Trek: The Next Generation*.

Figure 1. Elderly person opens mouth in imitation of AIBO opening its mouth. Photo courtesy of N. Edwards, A. Beck, P. Kahn, and B. Friedman.

A young adolescent male comes to greatly admire the android Data and begins to imitate him. The imitation starts innocently enough, but the boy soon captures more and more of Data's idiosyncratic mannerisms and personality. Is this scenario plausible? Consider that while demonstrating Sony's robotic dog AIBO to a group of elderly, Kahn and his colleagues caught a moment on camera where AIBO opened its mouth, and then an elderly person opened hers (see Figure 1). Thus here the second benchmark is: Will people come to imitate humanoid robots, and, if so, how will that compare to human–human imitation?

3. Intrinsic Moral Value

There are many practical reasons why people behave morally. If you hit another person, for example, that person may whack you back. Murder someone, and you will probably be caught and sent to jail. But underlying our moral judgments is something more basic, and moral, than simply practical considerations. Namely, psychological studies have shown that our moral judgments are in part structured by our care and value for people, both specific people in our lives, and people in the abstract (Kohlberg, 1984; Kahn, 1992; Turiel, 1983; Turiel, 1998). Although, in Western countries, such considerations often take shape in language around "human rights" and "freedoms" (Dworkin, 1978), they can and are found cross-culturally (Dworkin, 1978; Mei, 1972). Moreover, in recent years Kahn and his colleagues (Kahn, 1999) have shown that at times children and adults accord animals, and the larger natural world, intrinsic value. For example, in one study, a child argued that "Bears are like humans, they want to live freely… Fishes, they

want to live freely, just like we live freely… They have to live in freedom, because they don't like living in an environment where there is much pollution that they die every day" (p. 101). Here animals are accorded freedoms based on their own interests and desires.

The benchmark at hand, then, is: Will people accord humanoid robots intrinsic moral value (Kahn, Friedman, Perez-Granados, & Freier, 2006; Melson, Kahn, Beck, Friedman, Roberts, & Garrett, 2005)? Answering this question would help establish the moral underpinnings of human–robot interaction.

There is some initial evidence that in some ways people may accord robots intrinsic moral value. For example, Friedman, Kahn, and Hagman (2003) analysed 6,438 postings from three well-established online AIBO discussion forums. In their analysis, they provide some qualitative evidence of people who appear upset at the mistreatment of an AIBO and might well accord AIBO intrinsic moral value. For example, one member wrote, "I am working more and more away from home, and am never at home to play with him any more… he deserves more than that" (p. 277). In another instance, when an AIBO was thrown into the garbage on a live-action TV program, one member responded by saying: "I can't believe they'd do something like that?! Thats so awful and mean, that poor puppy…" Another member followed up: "WHAT!? They Actualy THREW AWAY aibo, as in the GARBAGE?!! That is outragious! That is so sick to me! Goes right up there with Putting puppies in a bag and than burying them! OHH I feel sick…" (p. 277). Thus one method is to garner people's judgments (either directly, as in asking questions; or indirectly, as emerges in discussion forum dialog) about whether robots have intrinsic moral value.

Yet part of the difficulty is that if you ask questions about robots, human interests are almost always implicated, and thus become a confound. For example, if I ask you, "Is it all right or not all right if I take a baseball bat and slug the humanoid?" you might respond, "It's not all right" — suggesting that you care about the humanoid's wellbeing. But upon probing, your reasoning might be entirely human-centered. For example, you might say: "It's not all right because I'll get in trouble with the robot's owner," or "because the humanoid is very expensive," or "because I'd be acting violently and that's not a good thing for me."

Thus, how is it possible to disentangle people's judgments about the intrinsic moral value of the robotic technology from other human-oriented concerns? One answer can be culled from a current study that investigates children's judgments about the intrinsic moral value of nature (Severson & Kahn, 2005). In this study, a new method was employed that set up a scenario where aliens came to an earth unpopulated by people, and the aliens caused harm to various natural constituents, such as animals and trees. Children were then interviewed about whether it was all right for the aliens to cause each of the natural constituents harm. Results

showed that children accorded nature intrinsic moral value at rates significantly higher than those found in comparison questions or in previous studies (Kahn, 1999). These results provide support for the alien methodology as a way to disentangle human considerations in assessing the intrinsic value of nature.

In turn, we have begun to explore whether a version of this method will work with a humanoid robot (using ATR's Robovie). Our inquiry focuses on isolation harm (e.g., is it all right or not all right for the aliens to stick the humanoid in a closet for a few years?), servitude (is it all right or not all right for the aliens to make the humanoid their personal servant?), ownership (is it all right or not all right for the aliens to buy and sell the humanoid?), and physical harm (is it all right or not all right for the aliens to crush the humanoid, like a used car?). If children believe that robots have intrinsic moral value, we would expect children to judge negatively the aliens' actions across these dimensions. Conversely, if children believe that robots do not have intrinsic moral value, we would expect children to accept the alien's actions. The one potential drawback of this method, however, is that it accords robots full autonomy insofar as they exist and function independent of humans. That may be fine if one seeks to examine whether people accord fully autonomous robots intrinsic moral value, but of less value in assessing judgments about robots as they currently exist or as they will exist as at least partial products of human creation.

Another method of approaching this benchmark of whether a humanoid robot has intrinsic moral value may involve the coordination of moral and personal judgments. What we have in mind here can be explicated in the following way. Consider a situation where a humanoid robot makes a moral claim on a person that conflicts with the person's own interests. For example, let's assume that a person (call him Daniel) has formed strong attachments to a humanoid, and Daniel believes that the humanoid has formed strong attachments to him. Let's then say that Daniel's house was recently burgled, and the humanoid tells Daniel: "I feel traumatized, and I'm scared staying home alone during the evenings. Another burglar might come. Daniel, would you please stay home with me during the evenings, at least for the next two weeks, while I have a chance to deal with my psychological distress?" The issue at hand is how Daniel coordinates the humanoid's moral claim with his (Daniel's) personal interests (the desire to spend some evenings away from one's home). The criterion question is whether the coordination is the same when the moral claim is made by a humanoid or by a human. For example, in the above situation, would people be equally inclined to stay home each evening for two weeks to help a humanoid as compared to a human friend and housemate?

4. Moral accountability

A defining feature of the moral life, and likely all legal systems, is that people of sound mind are held morally accountable for their actions. Indeed, that is partly why many people have difficulty accepting deterministic accounts of human life. For if behavior is fully determined by exogenous forces, such as contingencies of reinforcement or culture, or by endogenous forces, such as genes, then there appears no basis for holding people morally accountable for their actions. Granted, from a deterministic stance, you can still punish a perpetrator; but you cannot assign blame. For example, you would not be able to say to a man who steals money from the poor to support his lavish lifestyle, "You should not have done that." For the man could simply respond, "I'm not responsible for my behavior; I could not have done otherwise." And such responses seem to run roughshod over deeply held beliefs about human nature.

Accordingly, a benchmark is whether people will come to believe that humanoid robots are morally accountable for the behavior they cause (Friedman & Kahn, 1992). In our view, there would be two overarching categories of immoral behaviors to focus on, in particular. The first involves issues of unfairness or injustice. Imagine, for example, if a robotic daycare assistant unfairly distributes more treats to some children than to others? The criterion question is: Do people hold the humanoid itself morally responsible and blameworthy for unfair acts? The second involves the robot causing direct harm to people's welfare. In the moral-developmental literature (Turiel, 1998), three forms of welfare have been investigated extensively: physical (including injury, sickness, and death), material (including economic interests), and psychological (including comfort, peace, and mental health). The criterion question here is: Do people hold the humanoid itself morally responsible and blameworthy for acts that cause people direct harm?

In earlier research, Friedman and Millett (1995) explored this question in terms of whether undergraduate computer science majors believed that a computer system could be held morally accountable for acts that caused humans harm. For example, one scenario involved a computer system that administers medical radiation treatment and over-radiated a cancer patient because of a computer error. Results showed that 21% of the students interviewed consistently held computers morally responsible for such errors. Given that the stimulus (a computer system) mimicked only a small range of humanlike behavior, and that the participants were technologically savvy, there is good reason to believe that this benchmark — focused on judgments of moral accountability — will increasingly come into play as such systems take on increasingly sophisticated humanoid forms.

5. Privacy

Privacy refers to a claim, an entitlement, a right, or the ability of an individual (a) to determine what information about himself or herself can be communicated to others, and (b) to withdraw from society. The research literature suggests that children and adults need some privacy to develop a healthy sense of identity, to form attachments based on mutual trust, and to maintain the larger social fabric. The literature also shows that privacy in some form exists cross-culturally (Friedman & Kahn, 2003).

If humanoids (e.g., personal assistants for the home) become increasingly pervasive in human lives, and increasingly attain the ability to monitor and record personal information — and setting aside for the moment their ability to transmit that information — what is the effect on people's sense of privacy? A nascent issue along similar lines arises today with systems such as Google's Gmail. As analyzed by Friedman, Lin, and Miller (2006), each time a Gmail subscriber clicks on an email entry, the system retrieves the message and automatically scans the message for keywords provided by advertisers. Then the Google system selects and orders the advertisements to display on the subscriber's screen. In other words, a machine (not a person) "reads" subscribers' email. An open psychological question is whether people feel that this in some way compromises their privacy.

Or imagine a robot that moves around the floor of one's research lab, and chats with workers, and becomes their "friend," but also records the presence of individuals in the lab ("Hi Fred. I noticed yesterday you left early. Are you feeling okay?"), and through wireless connectivity keeps track of the flow and content of their email, and shares that information with other robots in the building or around town. Granted, if the robot is programmed to share that information with other humans, such as one's boss, then the robot has been turned partly into a surveillance system. But even if that capability is not designed into the robot, the benchmark is whether humanoids in and of themselves can encroach if not infringe on human privacy.

6. Reciprocity

Reciprocity is often viewed as being a central feature of the moral life. The "Golden Rule," for example, epitomizes one form reciprocity can take: "Do unto others as you would have them do unto you." Moreover, most moral-developmental theorists view reciprocal relationships as fundamental to the developmental process (Piaget, 1932/1969; Kohlberg, 1984; Turiel, 1998). For through reciprocal relationships, children take the perspective of others, recognize larger sets of problems that involve competing interests, and thereby seek to construct more adequate

solutions, more adequate in that the solutions address, for example, a larger set of competing interests. Note that, in this account, it would be difficult for children to develop morally if their primary relationships were ones where they were served by slaves. For there is little in that form of relationship that would require children to mutually "readjust" their interests and desires.

The benchmark then is, Can people engage substantively in reciprocal relationships with humanoids? The word "substantive" is important here, because it seems apparent that robots already do engage people in at least certain forms of reciprocal interactions. For example, if in meeting a robot, the robot extends its arm for a handshake, it is likely the human will respond in kind and shake the robot's hand (Figure 2). It is also possible to play air hockey with a robot. In Figure 2, for example, the person playing air hockey with the humanoid is anticipating the humanoid's next shot, and responding accordingly. Or, more formally, Kahn and colleagues (Kahn et al., 2006) analyzed 80 preschool children's reasoning about and behavior with AIBO (and a stuffed dog as a comparison artifact) over a 40-minute interactive session. In their behavioral analysis, they coded for six overarching

Figure 2. Reciprocal Interactions with Robots. Photo top: child playing "fetch" with Sony's robotic dog AIBO. Photo bottom left: ATR's Robovie initiates handshake. Photo bottom right: playing air hockey with robot. The air hockey research was performed by Darrin Bentivegna, and the robot was built by Sarcos (http://www.sarcos.com/) for ATR. (Photo credits, Value Sensitive Design Research Lab.)

behavioral categories: exploration, affection, apprehension, mistreatment, endowing animation, and reciprocity. Reciprocity was defined as the child's behavior not only responding to the artifact, but expecting the artifact to respond in kind based on the child's motioning behaviors, verbal directives, or offerings. For example, in Figure 2, AIBO is searching for the ball. The young boy observes AIBO's behavior and puts the ball in front of AIBO and says, "Kick it!" Based on an analysis of 2,360 coded behavioral interactions, Kahn et al. found that children engaged in significantly more attempts at reciprocity with AIBO (683 occurrences) than with the stuffed dog (180 occurrences). Indeed, reciprocity was by far the most frequently used category for interacting with AIBO compared to the next most frequently used category, affection (294 occurrences).

As robots gain an increasing constellation of humanlike features — as they increasingly have a persona ("personality"), adapt to social interactions, engage in "autonomous" (nondeterministic, but coherent) action, learn new behaviors, communicate, use natural cues, respond to emotions in humans, and self-organize (Minato et al, 2004; Fong, Nourbakhsh, & Dautenhahn, 2003) — it seems plausible to posit increasingly rich reciprocal interactions. One could imagine sometime in the future, for example, the following interaction between a humanoid robot (Jessie) playing a card game with a seven-year-old (Sam):

> Jessie: This will be really fun playing with you Sam; thanks for coming over.
> Sam: Sure, I was hoping you'd be free. Let's play five card draw.
> Jessie: Okay, but after that I was hoping we could play seven card draw; that's really my favorite.
> Sam: No way, I just want to play five card draw.
> Jessie: Well, gee, Sam, I don't want to play five card draw. Can't we just kind of trade off? Each take turns.
> Sam: I don't want to.
> Jessie: What do you think we should do? How can we solve this one?
> Sam: I just want you to do what I want.
> Jessie: No way, you can just go home, then.
> Sam: I don't want to… Well, I've got an idea, how about if there is a different game we both want to play. Do you like go fish?
> Jessie: Yeah, I love that game.
> Sam: Great! Deal 'em up!

Jessie's a robot. But can the robot's behavior set into motion the "opposition" of perspectives and desires that can occur in reciprocal interactions and which Piaget viewed as part of the mechanism (disequilibration) for the child's construction of morality?

The Oxford English Dictionary (2004) defines "reciprocal" as "[e]xisting on both sides; felt or shared by both parties; mutual." Setting aside the ontological

question of whether robots can actually feel or share, the human psychological issue remains. Thus, a criterion question that follows from this benchmark is whether people's reciprocal interactions with humanoids can be of the same form as with other people, or whether it takes on a strange hybrid unidirectional form, where the human is able ultimately to control or at least ignore the humanoid with social and moral impunity.

7. Conventionality

Social life includes social conventions: largely arbitrarily designated behaviors that promote the smooth functioning of social interactions (Turiel, 1983, 1998). For example, in some cultures acquaintances shake hands when meeting. Conventional practices are arbitrary in the sense that different practices, such as bowing or a namaste greeting, serve the same function equally well.

Over 100 published empirical studies have demonstrated that people distinguish conventional practices from moral behaviors (for reviews of the literature, see Helwig, Tisak, & Turiel, 1990; Smetana, 1995, 1997; Tisak, 1995; Turiel, 1983, 1998; Turiel, Killen, & Helwig, 1987). The distinction is based on two types of assessments. The first, *justifications*, refers to the reasons people provide for their normative judgments. Typically, conventional justifications focus on rules, common practice, and authority; in turn moral justifications focus on people's physical, material, and psychological welfare, and on claims to rights and justice. The second type of assessment, *criterion judgments*, refers to the criteria used to judge an act normatively. One criterion judgment comprises rule contingency (whether the normative judgment applies even if there was a rule that permitted the act). A second criterion judgment comprises generalizability (whether the judgment applies to other people with different customary practices in a different cultural location).

Here is an illustration of how this distinction between conventionality and morality plays out in interviews with children (Turiel, 1983; see Kahn, 1999, for an overview). Consider a school in the United States that requires children to call teachers by their surnames. Research typically shows the following pattern of results. The interviewer first asks a prescriptive question, often framed in terms of whether an act is all right or not all right to perform. In this scenario: "Is it all right or not all right to call teachers by their first names?" The student will answer "not all right," and often justify his or her evaluation based on an appeal to conventions ("Because that's the way we do things around here"). The interviewer then asks some version of a rule-contingency question: "Let's say that the principal of the school said that it is all right to call teachers by their first names, is it now all right or not all right to call teachers by their first names?" Now the student will answer

"all right." Then the interviewer asks some form of a generalizability question: "Let's say that the principal of a school in another country says that it is all right for students to call teachers by their first names; is it now all right or not all right for those students to call teachers by their first names?" The student will again answer "all right." Justifications for the two latter evaluations appeal to the relativity of conventions ("If people in authority decide to do things differently, they can").

In a second scenario, instead of asking about a conventional issue, the interviewer asks about a moral issue. Perhaps the interviewer asks about an event that involves unprovoked physical harm, such as when a bully starts a fight with a younger child. The student's evaluation for the first question will typically stay the same: "It is not all right for an older child to start a fight with the younger child." But the justifications differ, appealing to justice or human welfare ("It's not fair, because the younger child isn't doing anything wrong, and plus the older child could hurt him"). In addition, the student will say some version of the following: "It's not all right to start a fight even if the teacher and principal make a rule that says it is all right" (rule contingency); and "it's not all right even if another school in another country says it is, and allows it to occur" (generalizability). Justifications appeal to justice and human welfare, and are not contingent on personal interests, authority dictates, or common practice.

This distinction between conventional and moral judgments offers a benchmark for HRI. In a current study, for example, Freier (2007) set up an interaction between children and a 2-D conversational personified software agent. During the interaction, children witnessed the experimenter engaging in both a conventional violation with the software agent (the experimenter draws triangles instead of Xs and Os on a tic-tac-toe game board) and a moral violation (the experimenter insults the software agent by calling it names). Freier then interviewed the children, assessing how they construed the nature of the two types of violations, based on assessments of criterion judgments and justifications. Thus this study highlights a potentially important method to assess whether interaction with a humanlike robot is compelling in a humanlike way by investigating the distinction between conventionality and morality.

8. Creativity

Creativity involves imagining new and valuable ways to approach and solve problems. The literature on this topic supports the proposition that creativity exists to varying degrees in every person (John-Steiner, 2000; Sternberg, 2005, 2006; Sternberg & Lubart, 1991) and is "grounded in everyday abilities, such as conceptual thinking, perception, memory, and reflective self-criticism" (Boden, 2004, p. 1).

Creativity can take at least one of two forms. We propose the second form of creativity as a benchmark in human–robot interaction. Both forms can be illustrated in Nourbakhsh's (2006) curriculum on what he calls educational robotics. In this curriculum, children are offered the opportunity to design and construct robots from simple, readily available technologies (e.g., Telepresence Robotics Kit, http://www.terk.ri.cmu.edu/). In the first form, creativity emerges through a unidirectional process of acting on the robot as artifact. For example, a child may generate surprising and valuable solutions to problems encountered in the design and construction of robots. This first form of creativity — a unidirectional process of acting on an artifact — is foundational to human experience, characterizing, for example, much of an artist's or perhaps even physicist's experience of the world. The question at hand is whether people will treat humanlike robots as artifacts in the creative process.

In the second form, creativity emerges from interactions and collaborations with others. For example, Nourbakhsh (2006) found that team processes were valuable in generating innovative designs and implementations of robots, especially when a secondary goal was involved (e.g., robots navigating an obstacle course or creating a play with robot 'actors'). John-Steiner (2000) has written that in this form of creativity, "generative ideas emerge from joint thinking, from significant conversations, and from sustained, shared struggles to achieve new insights by partners in thought" (p. 3). In the following example (which we made up), consider children's interactions with each other in a pretend play setting. One child suggests a theme (e.g., "Let's play castles and princes"), and then the second child elaborates on this theme by introducing specific characters (e.g., "I'll be the prince and this [referring to a broom] will be my horse"). Then the first child offers additional ideas (e.g., "You ride your horse to the castle. This is the castle [pointing to the chair]. And I'm at the top of the castle [and the first child stands on the chair]. You're coming to find me so we can have a sword fight."), and so on as the story is elaborated. Social pretend play is thought to be more developmentally sophisticated (and it appears later) than solitary pretend play in that it has the added complexity of coordinating and integrating one another's contributions to form a more elaborate pretense (Fein, 1981).

Thus the question at hand here is not whether robots will become a medium to engender human–human creativity. Nor is the question whether robots can themselves be creative, or at least in their output. Rather, the question — the benchmark for human–robot interaction — is whether people will interact with robots as partners in a joint creative enterprise.

9. The authenticity of relation

In human–computer interaction (HCI), it is often assumed that the fundamental way people interact with computational systems is through "using" the system (Myers, Hollan, Cruz, et al., 1996). Such a view matches reasonably well with people's common use of language. Thus people talk, for example, of using computers to send email, using PDA's to keep track of their schedules, and using GPS devices to locate themselves on the planet. Yet even if the "use model" is appropriate for HCI, Freier (2006) has questioned whether the same model should be applied to interaction with humanlike robots.

To understand this critique, consider the use model in terms of our relationship with other people. Using other people can take at least one of two forms, both with pejorative connotations. In one form, a person controls another person by coercive means: "My great granddad was a slave during the 1800's, and the plantation owners used him however they wanted." In a second form, seemingly relational interactions become viewed as only self-serving: "She was only being nice to me [using me] because she wanted to meet my brother." "He was only using me to get ahead in the organization." "Now I understand that he was treating me [using me] as a sex object." In other words, many people are uncomfortable accepting a use model for human–human interaction, because they believe it fails to recognize the breadth and depth of people's social and moral interactions.

Thus, building on Freier (2006), we propose that, as in human–human interaction, in human–robot interaction a use model needs to be hierarchically integrated within an interactional theory. The reason is that humanlike robots will affect people in surprisingly rich ways, socially and morally; and the HRI field needs a corresponding theory to account for such experience. Elsewhere, Kahn (1999) has written about what such an interactional theory looks like (see also Kohlberg, 1969; Piaget, 1983, 1970; Turiel, 1983; Turiel & Davidson, 1986). In brief, interactional theory builds on constructivist psychological principles that delineate the developmental processes and mechanisms by which children — through interaction with a physical and social world — construct increasingly more adequate ways of understanding the world and acting on it. Such a theory also seeks to characterize fundamental categories of social and moral interaction. Indeed, most of the benchmarks proposed in this paper (e.g., autonomy, moral accountability, privacy, reciprocity, and conventionality) are themselves such categories, and thus begin to explicate what an interactional theory (as opposed to a use model) could look like for HRI.

The point we want to develop here is that perhaps even an interactional theory does not go far enough, and that interaction itself needs to be hierarchically integrated within what we will refer to as the authenticity of relation. What we have in

mind draws on Buber's (1970/1996) distinction between two fundamental types of relationships: "I-You" and "I-It." In an I-You relationship (also sometimes translated as "I-Thou"), an individual relates to another with his or her whole being, freely, fully in the present, unburdened by conceptual knowledge: "[t]he form that confronts me I cannot experience nor describe; I can only actualize it. And yet I see it, radiant in the splendor of the confrontation, far more clearly than all clarity of the experienced world" (p. 61). Buber continues, "The You encounters me. But I enter into a direct relationship to it. Thus the relationship is election and electing, passive and active at once… I require a You to become; becoming I, I say You" (p. 62). In contrast, in an I-It relationship an individual treats another individual much like an artifact: to be conceptualized, acted upon, and used.

Can an individual have an I-You relationship with a humanlike robot? It is not obvious how Buber would answer this question. According to Buber, any technology that increases the ability to use the world "generally involves a decrease in man's power to relate" (p. 92). Thus if a humanlike robot is understood only as a technological artifact, embedded in a use model of the world, then only an I-It relationship would seem possible. Yet what if humanlike robots are (at least from the psychological perspective) like people? Then individuals might be able to establish an I-You relationship with them.

Indeed, for Buber the I-You relationship can take partial forms, and such forms may more accurately come to reflect what is possible in human–robot interaction. One form can emerge with animals wherein, according to Buber, the You is latent: "In the perspective of our You-saying to animals, we may call this sphere the threshold of mutuality" (p. 173). Another form can emerge with plants. Buber writes:

> It is altogether different with those realms of nature which lack the spontaneity that we share with animals. It is part of our concept of the plant that it cannot react to our actions upon it, that it cannot "reply." Yet this does not mean that we meet with no reciprocity at all in this sphere. We find here not the deed of posture of an individual being but a reciprocity of being itself — a reciprocity that has nothing except being…Our habits of thought make it difficult for us to see that in such cases something is awakened by our attitude and flashes toward us from that which has being. What matters in this sphere is that we should do justice with an open mind to the actuality that opens up before us. This huge sphere that reaches from the stones to the stars I should like to designate as the pre-threshold, meaning the step that comes before the threshold. (p. 173)

Thus now the question becomes whether a humanlike robot can become a You at the "threshold of mutuality" like an animal, or at the "pre-threshold" of mutuality like a plant. Or perhaps a humanlike robot could become a You in some other way.

If in human–human interaction it is a challenge to know how to assess an I-You relationship, and it is, we would still maintain that this relationship, or something like it that focuses on the authenticity of relation, needs to be held out as a goal. It speaks to an essential, meaningful, and beautiful aspect of what is possible in human existence, and thus merits inclusion as a HRI benchmark.

Specifying the appropriate level of a psychological benchmark: Between abstraction and concretization

Toward the beginning of this paper, we defined psychological benchmarks as categories of interaction that capture conceptually fundamental aspects of human life, specified abstractly enough so as to resist their identity as a mere psychological instrument (e.g., as in a measurement scale), but capable of being translated into testable empirical propositions. We said that this definition should be understood as a first approximation, and that our definition would make more sense by considering the nine specific benchmarks offered in this paper. With those benchmarks behind us, we are now positioned to say more about one feature of psychological benchmarks that we view as particularly important: that of specifying their appropriate level between abstraction and concretization.

To do so, recall that one of the criterion judgments that Turiel (1983, 1998) uses to distinguish conventionality (our sixth benchmark) from morality is generalizability. This criterion has wide conceptual appeal. In courts of law, for example, it is generally agreed that like cases should be tried alike. (If a judge, for example, allows certain forms of evidence to be presented in her courtroom, one expects — from a moral perspective — that she should allow such forms of evidence to be presented in other similar cases.) Or if, from a moral perspective, a person claims that it is wrong for a country in times of war to torture prisoners, it is usually assumed that the person recognizes that the claim applies to their own country as well. That is, in moral philosophy, the universality of the claim (e.g., that it is wrong to torture prisoners) is part of what makes the claim a moral claim, as opposed to a narrowly self-serving interest.

Thus, Turiel extracted from a wide body of moral philosophical discourse a central feature — let us call it a benchmark — of the moral life: generalizability. In turn, to employ this benchmark, researchers have concretized it in specific ways. For example, some researchers have asked children the question: "Let's say that a child in China did X [the act under investigation, such as pushing another child off a swing]; would it be all right or not all right for a child in China to do X?" Here the researcher seeks to assess whether the child generalizes the normative judgment against X to a child in a similar situation but in a different culture. Other

researchers have assessed judgments of generalizability by first establishing a common practice within the different culture for the act under investigation. Such a question might read as follows: "Let's say that children in China did X, that's the way they do things there; in that case, would it be all right or not all right for a child in China to have done X?" This second form of the generalizability question poses a more stringent test of whether the moral judgment generalizes insofar as everyone within the different culture now engages in the act.

Our point is not to argue for a particular way of asking the generalizability question. Rather, we want to highlight that any such question is only a specific instantiation of the benchmark, and not the benchmark itself. In other words, Turiel's benchmark of generalizability operates — we think elegantly — at this intermediate level: conceptually abstract but with specificity within its sphere of influence; and able to be concretized into empirical assessments without being reduced to any specific form.

When examining our nine proposed benchmarks from this perspective, we have in our view achieved mixed success. Limitations can be seen, for example, with the benchmark of autonomy wherein we established conceptually its importance, but did not provide clear direction for how to assess it. Yet we did better with the benchmark of intrinsic moral value. Recall that we first framed the benchmark as a question: Will people accord humanlike robots intrinsic moral value? We then showed that the question posed methodological difficulties since human interests are almost always implicated in human–robot interaction, and it can be difficult to disentangle valuing a robot for its own sake from valuing a robot because of its effects on people's lives. We then proposed two general approaches toward disentangling these issues. One involved the "alien" methodology (where aliens mistreat robots, which thereby removes human interests from the social context). Another involved the coordination of personal and moral judgments: that is, when robots make moral claims that impinge on human personal interests, will people at times accept the validity of such robot claims?

In short, we are proposing that psychological benchmarks aim for as elegant and powerful a level as possible between abstraction and concretization, while providing ways to move forward with specific assessments. It is a difficult endeavor, and not fully achieved in this paper. But we hold out such specification as an ideal to strive toward.

Conclusion

Increasingly sophisticated humanoid robots will be designed and built, and in various ways integrated into our social lives. From the standpoint of human–robot

interaction, how do we measure success? In answering this question, we have suggested that the field could be well served by developing psychological benchmarks, and have offered nine contenders: autonomy, imitation, intrinsic moral value, moral accountability, privacy, reciprocity, conventionality, creativity, and authenticity of relation. As noted earlier, it is a tentative list, and in no way complete. One could well continue in this vein and offer benchmarks for emotion, attachment, cognition, and memory, for example. One could also try to establish benchmarks on the level of group interaction, as opposed to individual human–robot interaction. There are also important engineering benchmarks that need to be developed. That said, we believe our initial group of benchmarks make headway with the overall enterprise and help motivate why the enterprise itself is important.

How many benchmarks should be established in the field of HRI over the next decade? We are not sure. Perhaps around 25 to 35? If there are too few benchmarks, the field may pursue too narrow a vision of human–robot interaction. Too many benchmarks will likely indicate that the benchmarks themselves are not being characterized at a sufficiently high level of abstraction to capture robust, fundamental aspects of what it means to be a human.

To be clear, when we say "fundamental aspects of what it means to be a human," we do not mean that if a person does not have or has not fully realized these aspects that they do not exist as biological or psychological beings. A ruthless dictator may fare poorly on such benchmarks as intrinsic moral value, privacy, reciprocity, and authenticity of relation, but the dictator remains a person. In this way, our benchmarks are not framed as minimal requirements of personhood, but as teleological characterizations of what is possible in human existence.

Figure 3. Three Human Forms: Photo left: Japanese Sculpture. Photo bottom right: One version of ATR's Robovie. Photo top right: One of H. Ishiguro and colleagues' androids. (Photo credits: Value Sensitive Design Research Lab.)

To understand ourselves as a species is one of the profound undertakings of a lifetime. What we would like to suggest is that the study of human–robot interaction in general, and psychological benchmarks in particular, can provide a new method for such investigations. The idea is akin to that of comparative psychologists who have long studied animal behavior with the belief that by understanding our similarities and differences with other animal species, we discover more about our own (Povinelli, 2000; Tomasello, 2000). From the standpoint of HRI (e.g., see Figure 3), our basic move is that in investigating who we are as a species, and who we can become, we need not privilege the biological "platform."

The structure of this move dates back at least to the Turing test and is found in a good deal of research in artificial intelligence. But past attempts to understand humanity through investigations with computation have tended to focus narrowly on aspects of human cognition, and assumed that mental capacities could be abstracted from embodiment. Toward broadening the comparative move, Ishiguro and colleagues have recently proposed a new field, android science, that seeks to use androids to verify hypotheses for understanding humans (Ishiguro, 2004, 2005; Kanda, et al, 2004; MacDorman & Ishiguro, 2006).

Our current paper on psychological benchmarks builds on the ideas of android science; we seek to put into play the entirety of human psychology, extending not only into the realms of sociality but also morality. Consider, for example, our benchmark of moral accountability. Imagine a person walking in the woods, and a criminal jumps out from behind a rock and slices the person's throat. Most people would hold the criminal morally responsible for his actions. Now imagine the same situation, except a mountain lion jumps out from behind the rock and sinks his jaws into the person's throat and kills the person. Some people might want to hunt down the lion and kill it, so as to prevent the lion from harming more people. Nonetheless, most people would not hold the mountain lion morally accountable for its behavior; for moral accountability has traditionally been a uniquely human characteristic. But with our psychological benchmarks now in hand, we are able to ask whether in future years people will hold humanoids morally accountable for their behavior. If people do, if even partially, then aspects of the moral life which have till now been viewed as fundamentally coupled only with human experience may be viewed in a new way. And so it goes for each of the nine psychological benchmarks we have proposed in this paper. We may come to view each in a new way. Or not. The answers await further empirical investigations.

Either way, the psychological benchmarks serve their purpose, allowing us to build increasingly humanlike robots, and — in an increasingly technological world — helping us not to lose sight of what is possible, ethical, and beautiful in human life.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0325035. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Akiwa, Y., Sugi, Y., Ogata, T., & Sugano, S. (2004). Imitation based human–robot interaction: — Roles of joint attention and motion prediction. In *Proceedings of the 14th International Workshop on Robot and Human Interactive Communication* (pp. 283–288). New York: Association for Computing Machinery.

Alissandrakis, A., Nehaniv, C. L., Dautenhahn, K., & Saunders, J. (2006). Evaluation of robot imitation attempts: Comparison of the system's and the human's perspectives. In *Proceedings of the 1st Annual Conference on Human–Robot Interaction* (pp. 134–141). New York: Association for Computing Machinery.

Aylett, R. (2002). *Robots: Bringing intelligent machines to life?* Hauppauge, NY: Barron.

Baldwin, J. M. (1973). *Social and ethical interpretations in mental development.* New York: Arno. (Original work published 1897).

Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2005, July 22–27). A cross-cultural study on attitudes towards robots. *Proceedings 11th International Conference on Human–Computer Interaction (HCI International 2005)*, Las Vegas, USA.

Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd edition). New York: Routledge.

Breazeal, C. L. (2002). *Designing sociable robots: Intelligent robotics and autonomous agents*. Cambridge, MA: MIT Press.

Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., & Blumberg, B. (2005). Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life, 11*, 31-62.

Breazeal, C., & Scassellati, B. (2002). Robots that imitate humans. *Trends in Cognitive Sciences*, *6*, 481–487.

Buber, M. (1996). *I and thou* (M. Kaufmann, Trans.). New York: Touchstone. (Original work published 1970)

Buchsbaum, D., Blumberg, B., Breazeal, C., & Meltzoff, A. N. (2005). A simulation-theory inspired social learning system for interactive characters. In *Proceedings of the 14th International Workshop on Robot and Human Interactive Communication* (pp. 85–90). Piscataway, NJ: IEEE.

Dautenhahn, K. (2003). Roles of robots in human society — Implications from research in autism therapy. *Robotica, 21,* 443–452.

Dautenhahn, K., & Nehaniv, C. L. (Eds.). (2002). *Imitation in animals and artifacts.* Cambridge, MA: MIT Press.

Dawkins, R. (1976). *The selfish gene*. New York: Oxford University Press.

Dworkin, R. (1978). *Taking rights seriously*. Cambridge: Harvard University Press.

Fein, G. G. (1981). Pretend play in childhood: An integrative review. *Child Development, 52,* 1095–1118.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42,* 143–166.

Freier, N. G. (2006). Towards an interactional model of children's relationships to personified adaptive systems. *Proceedings of the 5th International Conference on Cognitive Science (ICCS 2006)* (pp. 91–92). Vancouver, B.C., Canada.

Freier, N. G. (2007). Children distinguish conventional from moral violations in interactions with a personified agent. In *Extended Abstracts of the Conference on Human Factors in Computing (CHI '07).* (pp. 2195–2200). New York: Association for Computing Machinery.

Friedman, B., & Kahn, P. H., Jr. (1992). Human agency and responsible computing: Implications for computer system design. *Journal of Systems Software, 17,* 7–14.

Friedman, B., & Kahn, P. H., Jr. (2003). Human values, ethics, and design. In J. A. Jacko & A. Sears (Eds.), *The Human–computer interaction handbook* (pp. 1177–1201). Mahwah, NJ: Erlbaum.

Friedman, B., Kahn, P.H., Jr., & Hagman, J. (2003). Hardware companions? What online AIBO discussion forums reveal about the human-robotic relationship. *Proceedings of the CHI 2003 Conference on Human Factors in Computing Systems* (pp. 273-280). New York: Association for Computing Machinery.

Friedman, B., Lin, P., & Miller, J. K. (2005). Informed consent by design. In L. Cranor & S. Garfinkel (Eds.), *Designing Secure Systems that People Can Use* (pp. 495–521). Cambridge, MA: O'Reilly and Associates.

Friedman, B., & Millet, L. (1995). "It's the computer's fault": Reasoning about computers as moral agents. In *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 226–227). New York: Association for Computing Machinery.

Gopnik, A., & Meltzoff, A. N. (1998). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Helwig, C. C., Tisak, M., & Turiel, E. (1990). Children's social reasoning in context. *Child Development, 61*, 2068–2078.

Hofstadter, D. R., & Dennett, D. C. (Eds.). (1981). *The mind's I*. New York: Basic Books.

Ishiguro, H. (2004). Toward interactive humanoid robots: A constructive approach to developing intelligent robots. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems* (AAMAS 2004) (pp. 621–622). New York: Association for Computing Machinery.

Ishiguro, H. (2005). Android science: Toward a new cross-disciplinary framework. In *Cogsci–2005 workshop: Toward social mechanisms of android sciences* (pp. 1–6). Stresa, Italy.

John-Steiner, V. (2000). *Creative Collaboration.* New York: Oxford University Press.

Kahn, P. H., Jr. (1992). Children's obligatory and discretionary moral judgments. *Child Development, 63,* 416–430.

Kahn, P. H., Jr. (1999). *The human relationship with nature: Development and culture*. Cambridge, MA: MIT Press.

Kahn, P. H., Jr., Freier, N. G., Friedman, B., Severson, R. L. & Feldman, E. (2004). Social and moral relationships with robotic others? *Proceedings of the 13th International Workshop on Robot and Human Interactive Communication (RO-MAN '04)* (pp. 545–550). Piscataway, NJ: IEEE.

Kahn, P. H., Jr., Friedman, B., Perez-Granados, D. R., & Freier, N. G. (2006). Robotic pets in the lives of preschool children. *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, *7*, 405–436.

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human Computer Interaction*, *19*, 61–84.

Kanda, T., Ishiguro, H., Imai, M., & Ono, T. (2004). Development and evaluation of interactive humanoid robots. In *Proceedings of the IEEE (Special issue on Human interactive robot for psychological enrichment), 92*, 1839–1850.

Kaplan, F. (2001). Artificial attachment: Will a robot ever pass Ainsworth's strange situation test? In S. Hashimoto (Ed.), *Proceedings of Humanoids 2001: IEEE-RAS International Conference on Humanoid Robots* (pp. 99–106). Piscataway, NJ: IEEE.

Kiesler, S., & Goetz, J. (2002). Mental models of robotic assistants. In *Extended Abstracts of the Conference on Human Factors in Computing Systems* (CHI '02) (pp. 576–577). New York: Association for Computing Machinery.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). New York: Rand McNally.

Kohlberg, L. (1984). *Essays in moral development: Vol. II. The psychology of moral development.* San Francisco: Harper & Row.

MacDorman, K. F. (2005). Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it? *CogSci–2005 Workshop: Toward Social Mechanisms of Android Science* (pp. 106–118). Stresa, Italy.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, *7*, 297–337.

Mei, Y. P. (1972). Mo Tzu. In P. Edwards (Ed.), *The Encyclopedia of Philosophy, Vol. 5.* (pp. 409–410). New York: Macmillan.

Melson, G. F., Kahn, P. H., Jr., Beck, A. M., Friedman, B., Roberts, T., & Garrett, E. (2005). Robots as dogs? — Children's interactions with the robotic dog AIBO and a live Australian Shepherd. In *Extended Abstracts of the Conference on Human Factors in Computing Systems* (CHI '05) (pp. 1649–1652). New York: Association for Computing Machinery.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*, 838–850.

Minato, T., Shimada, M., Ishiguro, H., & Itakura, S. (2004). Development of an android robot for studying human–robot interaction. In *Proceedings of the 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems* (pp. 424–434). New York: Association for Computing Machinery.

Myers, B., Hollan, J., Cruz, I., Bryson, S. Bulterman, D., Catarci, T., Citrin, W., Glinert, E., Grudin, J., & Ioannidis, Y. (1996). Strategic directions in human–computer interaction. *ACM Computing Surveys*, *28*, 794–809.

Nourbakhsh, R. I. (2006). A roadmap for technology literacy and a vehicle for getting there: Educational robotics and the TeRK project. Plenary presented at *The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '06),* September 6–8, University of Hertfordshire, Hatfield, Hertfordshire, UK.

Oxford English Dictionary. (2004). Oxford University Press, Oxford, England. Retrieved May 24, 2004 from http://dictionary.oed.com/.

Piaget, J. (1969). *The moral judgment of the child*. Glencoe, IL: Free Press. (Original work published 1932).

Piaget, J. (1970). *Structuralism*. New York: Harper and Row.

Piaget, J. (1983). Piaget's theory. In W. Kessen (Ed.), P. H. Mussen (Series Ed.), *Handbook of child psychology: Vol. 1. History, theory, and methods* (4th ed., pp. 103–128). New York: Wiley.

Povinelli, D. J. (2000). *Folk physics for apes: The chimpanzee's theory of how the world works*. New York: Oxford University Press.

Searle, J. R. (1990). Is the brain's mind a computer program? *Scientific American, 262, 1,* 26–31.

Severson, R. L. & Kahn, P. H., Jr. (2005, April). *Social and moral judgments about pesticides and the natural environment: A developmental study with farm worker children.* Paper presented at the biennial meeting of the Society for Research in Child Development, Atlanta, GA.

Skinner, B. F. (1974). *About behaviorism*. New York: Knopf.

Smetana, J. G. (1995). Morality in context: Abstractions, ambiguities and applications. In R. Vasta (Ed.), *Annals of Child Development* (Vol. 10, pp. 83–130). London: Jessica Kingsley.

Smetana, J. (1997). Parenting and the development of social knowledge reconceptualized: A social domain analysis. In J. E. Grusec & L. Kuczynski (Eds.), *Handbook of parenting and the transmission of values* (pp. 162–192). New York: Wiley.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human–robot interaction. In *Proceedings of the 1st Annual Conference on Human–Robot Interaction* (HRI '06) (pp. 33–40). New York: Association for Computing Machinery.

Sternberg, R. J. (2005). Creativity or creativities? In E. A. Edmonds & L. Candy (Eds.), Special Issue on Computer support for creativity. *International Journal Human–Computer Studies, 63,* 370–382.

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal, 18,* 87–98.

Sternberg, R. J. & Lubart, T. I. (1991). An investment theory of creativity and its development. *Human Development, 34,* 1–31.

Tisak, M. S. (1995). Domains of social reasoning and beyond. In R. Vasta (Ed.), *Annals of Child Development* (Vol. 11, pp. 95–130). London: Jessica Kingsley.

Tomasello, M. (2000). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Turiel, E. (1983). *The development of social knowledge*. Cambridge, England: Cambridge University Press.

Turiel, E. (1998). Moral development. In W. Damon (Ed.), *Handbook of child psychology*. (5th ed.). Vol. 3: N. Eisenberg (Ed.), *Social, emotional, and personality development* (pp. 863–932). New York: Wiley.

Turiel, E. & Davidson, P. (1986). Heterogeneity, inconsistency, and asynchrony in the development of cognitive structures. In I. Levin (Ed.), *Stage and structure: Reopening the debate* (pp. 106–143). Norwood, NJ: Ablex.

Turiel, E., Killen, M., & Helwig, C. C. (1987). Morality: Its structure, functions and vagaries. In J. Kagan and S. Lamb (Eds.), *The emergence of morality in young children* (pp. 155–244). Chicago: University of Chicago Press.

Yamamoto, D., Doi, M., Matsuhira, N., Ueda, H., & Kidode, M. (2004). Behavior fusion in a robotic interface for practicality and familiarity: Approach by simultaneous imitations. In *Proceedings of the 14th International Workshop on Robot and Human Interactive Communication (HRI '04)*, (pp. 114–119). New York: Association for Computing Machinery.

Authors' addresses

Peter H. Kahn, Jr.
Department of Psychology
Box 351525
University of Washington
Seattle
WA 98195–1525 USA

pkahn@u.washington.edu

Batya Friedman
The Information School
University of Washington
Box 352840
Seattle
WA 98195–2840 USA

batya@u.washington.edu

Takayuki Kanda
Intelligent Robotics and Communication Laboratory
ATR
2–2–2 Hikaridai Keihanna Science City
Kyoto 619–0288 Japan

kanda@atr.jp

Nathan G. Freier
The Information School
Box 352840
University of Washington
Seattle
WA 98195–2840 USA

nfreier@u.washington.edu

Hiroshi Ishiguro
Department of Adaptive Machine Systems
Osaka University
2–1 Yamadaoka, Suita
Osaka 565–0871 Japan

ishiguro@ams.eng.osaka-u.ac.jp

Rachel L. Severson
Department of Psychology
Box 351525
University of Washington
Seattle
WA 98195–1525 USA

raches@u.washington.edu

Jessica Miller
Department of Computer Science and Engineering
Box 352350
University of Washington
Seattle
WA 98195–2350 USA

jessica@cs.washington.edu

About the authors

Peter H. Kahn, Jr. is Associate Professor in the Department of Psychology and Adjunct Associate Professor in the Information School at the University of Washington. He received his Ph.D. from the University of California, Berkeley in 1988. His research interests include human–robot interaction, the psychological effects when technologies mediate the human experience of nature, and value sensitive design.

Hiroshi Ishiguro is Professor in the Department of Adaptive Machine Systems at Osaka University and Visiting Group Leader at ATR, Japan. He received his B.Eng. and M.Eng. in Computer Science from Yamanashi University, Japan in 1986 and 1988, respectively, and his D.Eng in Systems Engineering from Osaka University, Japan in 1991. His research interests include distributed vision systems, robotics, and android science.

Batya Friedman is Professor in the Information School and Adjunct Professor in the Department of Computer Science and Engineering at the University of Washington where she co-directs the Value Sensitive Design Research Laboratory. She received her BA in 1979 and Ph.D. in 1988 from the University of California, Berkeley. Her research interests include value sensitive design, physical and cultural adaptation to information technologies, design methods, and human–robot interaction.

Takayuki Kanda is Senior Researcher at ATR Intelligent Robotics and Communication Laboratories, Japan. He received his B.Eng, M.Eng, and Ph.D. degrees in Computer Science from Kyoto University, Japan, in 1998, 2000, and 2003, respectively. His current research interests include intelligent robotics and human–robot interaction.

Nathan G. Freier is a doctoral student in the Information School at the University of Washington. He received his B.S. with Distinction in Computer Science and his B.A. in Comparative History of Ideas, both in 2000, from the University of Washington. His research interests include human–robot interaction, computer-mediated communication, and virtual environments.

Rachel L. Severson is a doctoral student in Developmental Psychology at the University of Washington. Her research interests focus on social and moral development, in particular investigating the role of pretense and imagination, theory of mind, and psychological conceptions of natural (e.g., animal) and computational (e.g., robotic) entities.

Jessica Miller is a doctoral student in Computer Science and Engineering at the University of Washington. Her research interests focus on how emerging technologies (e.g. humanoid robots, wearable devices, ubiquitous computing) enhance or degrade the quality of social relationships. Her most recent work has focused on how value sensitive design can aid the design, implementation, and adoption of a groupware tool she helped design and build for research engineers.