

# What is a Human? – Toward Psychological Benchmarks in the Field of Human-Robot Interaction

Peter H. Kahn, Jr., Hiroshi Ishiguro, Batya Friedman, and Takayuki Kanda

**Abstract—** In this paper, we move toward offering psychological benchmarks by which to measure success in building increasingly human-like robots. By psychological benchmarks we mean 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, but capable of being translated into testable empirical propositions. Six possible benchmarks are considered: autonomy, imitation, intrinsic moral value, moral accountability, privacy, and reciprocity. 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.

## I. INTRODUCTION

In various subfields within 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](http://www.tpc.org)), which comprises a data set with which to populate one's database, and a set of queries to run on the data base. Then one can compare the performance of one's system to those of the wider community. But in the field of human-robot interaction, if one of the goals is to build increasingly human-like 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 distinguish between ontological and psychological claims about such humanoids. Then we offer six 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

Manuscript received March 13, 2006. 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.

P. H. Kahn, Jr., is with the Department of Psychology, Box 351525, University of Washington, Seattle, WA 98195 USA (phone: 206-616-9395; fax: 206-685-3157; e-mail: [pkahn@u.washington.edu](mailto:pkahn@u.washington.edu)).

H. Ishiguro is with both the Department of Adaptive Systems, Osaka University, 2-1 Yamadaoka, Suita City, Osaka, Japan and the Intelligent Robotics and Communication Laboratory, ATR, 2-2-2 Hikaridai Keihanna Science City, Kyoto, Japan (e-mail: [ishiguro@atr.jp](mailto:ishiguro@atr.jp)).

B. Friedman is with the Information School, Box 352840, University of Washington, Seattle, WA 98195 USA (e-mail: [batya@u.washington.edu](mailto:batya@u.washington.edu)).

T. Kanda is with the Intelligent Robotics and Communication Laboratory, ATR, 2-2-2 Hikaridai Keihanna Science City, Kyoto, Japan (e-mail: [kanda@atr.jp](mailto:kanda@atr.jp)).

what constitutes essential features of being human.

## II. WHY BUILD HUMANOID ROBOTS?

We want to acknowledge that there are some good reasons *not* to have the goal to build human-like robots. One reason, of course, is that in many forms of human-robot interaction there is nothing gained functionally by going the humanoid route, as with, for example, industrial robots on an assembly line. There are also presumably numerous contexts where the human-like form may work against optimal human-robot interaction. For example, an elderly person being helped to the bathroom by a robotic assistant may not want the robot to have a human-like face so as to not be “looked” at by the robot during such personal moments. It is also true that given current technical limitations, humans may negatively evaluate a robot that has too much of the appearance but too little of the behavioral repertoire of a human, part of a phenomenon which is sometimes referred to in the literature as the “Uncanny Valley” [1, 2].

That said, and not to belabor the point, there are equally good reasons to aim to build human-like robots. Functionally, for example, human-robot communication will presumably be optimized in many contexts if the robot conforms to human-like appearance and behavior, rather than asking humans to conform to a computational system [3, 4, 5, 6]. It is also possible that psychological benefits could accrue if humans kept “company” with robotic others [7]. 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 16<sup>th</sup> century.

## III. 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 [8] 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, psychological, focuses on what people attribute to the fully human-like humanoid. The strong psychological claim is that people would conceive of the humanoid as human. The weak psychological claim is that

people would conceive of the humanoid 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 ontology. In Cases 2 and 3, people's psychological beliefs and actions would not be in accord with the correct ontological status of the robot.

Our point here is to be very clear that there is a distinction between claims regarding the ontological status of humanoid robots, and the psychological stance people bring toward them. It is important because a large amount of debate in the field of cognitive science and artificial intelligence has been centered on ontological questions, such as: are computers as we can conceive of them today in material and structure capable of becoming conscious? [9]. 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 our view, we think it likely in terms of societal functioning and wellbeing that the psychological question is at least as important as the ontological question.

#### IV. TOWARD PSYCHOLOGICAL BENCHMARKS

The issue at hand then becomes, psychologically speaking how do we measure success in building human-like robots? One approach might be to take findings from the psychological scientific disciplines, and seek to replicate them in human-robotic 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-robotic interaction. But, aside from good intuitions and luck, on what bases does one choose which studies to replicate? Indeed, given that human-robotic 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 moving forward over these next few years toward 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 done to date on examining people's human-like responses to robots [e.g., 1, 10, 11, 12, 13, 14] and on common metrics for task-oriented human-robot interaction [15], we know of no literature in the field that has taken such a direct approach toward establishing psychological benchmarks.

#### V. SIX PSYCHOLOGICAL BENCHMARKS TO CONSIDER

With the above working definition in hand, we offer the following six psychological benchmarks. Some of the benchmarks are characterized with greater specificity than others, given 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 differently. 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.

##### A. *Autonomy*

A debated issue in the social sciences is of whether humans themselves are autonomous. Psychological behaviorists [16], 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 needs be postulated. Dawkins [17] 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 becomes moral. For example, in his early work, Piaget [18] 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; while 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 [19] proposed that only by the latter stages of moral development (occurring in adolescence, if ever) does moral thinking differentiate from fear of punished and personal interests (stages 1 and 2) as well as conventional expectations and obedience to social systems (stages 3 and 4) to be characterized as autonomous

(stages 5 and 6).

Now, 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 [20] 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(ism), 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? This question seems to us of a high-order "benchmark" level. And the answer to the question 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.

### B. Imitation

Evidence suggests that neonates engage in rudimentary imitation, such as imitating the facial gestures of adults (see Figure 1). Then, through development, it is commonly agreed that imitation occurs across an increasingly wide sphere of contexts and on increasingly complex and abstract phenomena. Disputed in the field, however, is to what extent imitation can be categorized as rote, virtually "mindless" activity, or whether imitation needs to be conceptualized as a highly active, constructive process [21, 22].

A partial account of the constructive process can be drawn from the work of James Mark Baldwin. According to Baldwin [23], there are three circular processes (which Baldwin also conceptualizes as stages) 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



Fig. 1. Imitation of facial gestures by human neonates. Photo courtesy of A. Meltzoff. From Meltzoff, A. N., and Moore, M. K. Imitation of facial and manual gestures by human neonates. *Science*, 198, 4312 (Oct. 1977), 75-78.



Fig. 2. Elderly person opens mouth in imitation of AIBO opening its mouth. (The woman's first encounter with a robotic dog.) Photo courtesy of N. Edwards, A. Beck, P. Kahn, and B. Friedman.

plane is also a real invention...He makes it; he gets it for himself by his own action; he achieves, invents it" (p. 106).

One early benchmark in the field of HRI has been the Turing Test, originally known as the "imitation game" [24]. A computer system passes the Turing Test when a person ("the interrogator") cannot systematically distinguish between the computer system and a human being on the basis of their respective responses to questions posed by the interrogator.

It seems likely that humanoid robots will be increasingly designed to imitate people, not only using language-based interfaces, but through appearance and an increasing range of human-like behaviors [25, 26, 27, 28, 29, 30]. In turn, there are two additional imitation benchmarks. One is whether people will come to believe that humanoid robots imitate in a passive or active manner. Of course, as with the benchmark of autonomy, one will need clear assessments, as well, of whether people believe that other people imitate in a passive or active manner.

A second is perhaps even more interesting, and can be motivated by a fictional episode from the television program *Star Trek: The Next Generation*. A young adolescent male comes to greatly admire the character Data, who is an android. As adolescents are sometimes wont to do with those they admire, the adolescent begins to imitate Data. The imitation starts innocently enough, but as the story progresses the boy captures more and more of Data's idiosyncratic mannerisms (such as Data's unique speech pattern and head motion) and personality. Is this scenario plausible? Consider that while demonstrating 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 2). Thus here the second benchmark is: Will people come to imitate humanoid robots, and, if so, how will that compare to human-human imitation?

### C. Intrinsic Moral Value

There are many practical reasons why people behave morally. If you hit another person, for example, that person can 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 [19, 31, 32, 33]. Although, in Western countries, such considerations often take shape in language around "human rights" and "freedoms" [34], they can and are found cross-culturally [33, 35]. Moreover, in recent years Kahn and his colleagues [20] have shown that at times children (and people in general) also 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 the animals' own interests and desires ("they don't like" dying).

The benchmark at hand, then, is: Will people accord humanoid robots intrinsic moral value [36, 37]? Answering this question would go some distance toward establishing the moral underpinnings of human-robot interaction.

This question, however, is not so easy to answer. 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 "it's not all right because the humanoid is very expensive," or "it's not all right because I'll be acting violently and that's not a good thing for me."

In a current developmental study that investigates children's judgments about the intrinsic moral value of nature [38], progress methodologically has been made dealing with this difficulty by setting up a scenario where Aliens come to earth unpopulated by people, and then probing children's reasoning about whether or not the Aliens can harm various natural constituents, such as animals and trees. We are beginning to explore whether a version of this method will work with a humanoid robot, focusing on specific dimensions such as *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?).

Another way to gain traction methodologically on 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, where the claim conflicts with the person's personal 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 burglarized, 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 personal friend and housemate?

### D. 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 part of the reason 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.” Or, “You could have behaved otherwise.” 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 [39]. 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 others? The criterion question is: Do people hold the humanoid, itself, morally responsible and blameworthy for unfair or unjust acts? The second involves the robot causing direct harms to people’s welfare. In the moral-developmental literature [33], 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 [40] 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 due to a computer error over-radiated a cancer patient. 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 human-like 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.

#### *E. Privacy*

Privacy refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others. 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. (See Friedman & Kahn [41], for an introduction to the conceptual and empirical literature

on privacy.)

The issue then at hand is the following: 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 [42], each time a Gmail subscriber clicks on an email entry, the system retrieves the message and scans the message for keywords provided earlier by advertisers. Then the Google system selects and orders the advertisements to display on the subscriber’s screen. In other words, a machine “reads” subscribers’ email. An open current psychological question is whether people thereby feel that their privacy is in some way compromised.

These questions become ever more compelling when the scenario changes to living or working with humanoids. Imagine a robot, for example, 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, 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.

#### *F. Reciprocity*

Reciprocity is often viewed to be 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 [18, 19, 33]. 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 “adjust” their interests and desires.

The benchmark then is, Can people engage substantively in reciprocal relationships with humanoids? The word “substantive” is important here. For 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 3). It is also possible to play air hockey with a robot. In Figure 3, for example, the person is anticipating the robot's next shot, and responding accordingly. Or, more formally, Kahn and colleagues [36] analyzed 80 preschool children's reasoning about and behavior with Sony's robotic dog AIBO (and a stuffed dog as a comparison artifact) over a 40-minute interactive session. In their behavioral analysis, they coded for 6 overarching behavioral categories: exploration, affection, 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 3, 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 (683 occurrences of reciprocity compared to 294 occurrences of affection).

As robots gain an increasing constellation of human-like features – as they increasingly have a persona ("personality"), adapt to social interactions, engage in "autonomous" (non deterministic, but coherent) action, learn new behaviors, communicate, use natural cues, respond to emotions in humans, and self-organize [6, 43] – then it seems to us 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 7-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 5 card draw.

Jessie: Okay, but after that I was hoping we could play 7 card draw; that's really my favorite.

Sam: No way, I just want to play 5 card draw.

Jessie: Well, gee, Sam, I don't want to play 5 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.

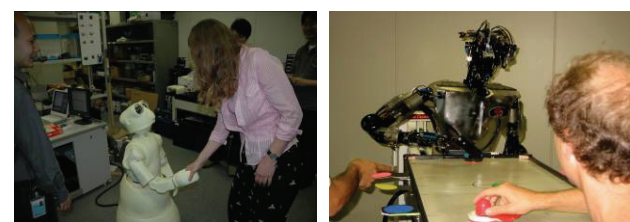


Fig. 3. Reciprocal Interactions with Robots. Photo top: child playing "fetch" with Sony's robotic dog AIBO (photo credit, Value Sensitive Design Research Lab). Photo bottom left: ATR's Robovie initiates handshake (photo credit, P. Kahn). Photo bottom right: playing air hockey with robot (photo credit, N. Freier). The air hockey research was performed by Darrin Bentivegna, and the robot was built by Sarcos (<http://www.sarcos.com/>) for ATR.

Sam: I don't want to go home.

Jessie: I don't know then....

Sam: 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 (disequilibrium) for the child's construction of morality?

The Oxford English Dictionary [44] 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.

## VI. CONCLUSION

Increasingly sophisticated humanoid robots will be designed and built, and in various ways integrated into our



Fig. 4. Three Human Forms: Photo left: Japanese Sculpture. Photo bottom right: One version of ATR's Robovie. Photo top right: One of H. Ishiguro's and colleagues androids. (Photo credits: P. Kahn.)

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 six contenders: autonomy, imitation, intrinsic moral value, moral accountability, privacy, and reciprocity. 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, memory, and creativity, 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're not sure. Perhaps around 20 to 30? Too few benchmarks and likely enough the field will 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 high-enough level of abstraction to capture robust, fundamental aspects of human interaction.

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 to other animal species, we discover more about our own. For example, based on his investigations of chimpanzee cognition, Povinelli [45] argues that humans have a unique capacity to generate concepts related to perceptually non-obvious phenomena, and that this capacity may be "one of the critical 'triggers' that unleashed human populations into

nearly every ecogeographic zone on the planet appropriately 200,000 years ago, while the species of great apes remained restricted to the tropics and neotropics" (p. 339). Along similar lines, Tomasello [46] has suggested that this same human capacity lays the foundation for cumulative cultural learning – a "ratchet effect" – that precipitates modern-day human achievements. From the standpoint of HRI (e.g., see Figure 4), the new move is that in investigating who we are as a species, and who we can become, we need not privilege the biological "platform."

#### ACKNOWLEDGMENT

Our thanks to Nathan G. Freier, Jessica K. Miller, and Rachel L. Severson, doctoral students at the University of Washington, for long-standing discussions on some of the central ideas in this paper.

#### REFERENCES

- [1] K. Dautenhahn, "Roles of robots in human society – Implications from research in autism therapy." *Robotica*, vol. 21, no. 4, pp. 443-452, Aug. 2003.
- [2] M. Mori, *The Buddha in the Robot*. North Clarendon, VT: Tuttle, 1982.
- [3] H. Ishiguro, "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*, New York, 2004, pp. 621-622.
- [4] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial." *Human Computer Interaction*, vol. 19, no. 1-2, pp. 61-84, 2004.
- [5] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Development and evaluation of interactive humanoid robots." *Proceedings of the IEEE (Special Issue on Human Interactive Robots for Psychological Enrichment)*, vol. 92, no. 11, pp. 1839-1850, Nov. 2004.
- [6] T. Minato, M. Shimada, H. Ishiguro, and S. Itakura, "Development of an android robot for studying human-robot interaction," in *Proceedings of the 17<sup>th</sup> International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, Ottawa, Ontario, Canada, 2004, pp. 424-434.
- [7] P. H. Kahn, Jr., N. G. Freier, B. Friedman, R. L. Severson, and E. Feldman, "Social and moral relationships with robotic others?" in *Proceedings of the 13<sup>th</sup> International Workshop on Robot and Human Interactive Communication*, Kurashiki, Okayama, Japan, 2004, pp. 545-550.
- [8] J. R. Searle, "Is the brain's mind a computer program?" *Scientific American*, vol. 262, no. 1, pp. 26-31, Jan. 1990.
- [9] D. R. Hofstadter, and D. C. Dennett, Eds., *The Mind's I*. New York, NY: Basic Books, 1981.
- [10] R. Aylett, *Robots: Bringing Intelligent Machines to Life?* Hauppauge, NY: Barron, 2002.
- [11] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "A cross-cultural study on attitudes towards robots," presented at the 11th International Conference on Human-Computer Interaction, Las Vegas, NV, USA, July 22-27, 2005.
- [12] C. L. Breazeal, *Designing Sociable Robots: Intelligent Robotics and Autonomous Agents*. Cambridge, MA: The MIT Press, 2002.
- [13] F. Kaplan, "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*, Tokyo, Japan, 2001, pp. 99-106.

- [14] S. Kiesler, and J. Goetz, "Mental models of robotic assistants," in *Extended Abstracts of the Conference on Human Factors in Computing Systems*, Minneapolis, MN, USA, 2002, pp. 576-577.
- [15] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, "Common metrics for human-robot interaction," in *Proceedings of the 1<sup>st</sup> Annual Conference on Human-Robot Interaction*, Salt Lake City, UT, 2006, pp. 33-40.
- [16] B. F. Skinner, *About Behaviorism*. New York, NY: Knopf, 1974.
- [17] R. Dawkins, *The Selfish Gene*. New York, NY: Oxford University Press, 1976.
- [18] J. Piaget, *The Moral Judgment of the Child*. Glencoe, IL: Free Press, 1969. (Original work published 1932.)
- [19] L. Kohlberg, *Essays in Moral Development: Vol. II. The Psychology of Moral Development*. San Francisco, CA: Harper & Row, 1984.
- [20] P. H. Kahn, Jr., *The Human Relationship with Nature: Development and Culture*. Cambridge, MA: The MIT Press, 1999.
- [21] A. Gopnik, and A. N. Meltzoff, *Words, Thoughts, and Theories*. Cambridge, MA: The MIT Press, 1998.
- [22] A. N. Meltzoff, "Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children," *Developmental Psychology*, vol. 31, no. 5, pp. 838-850, Sep. 1995.
- [23] J. M. Baldwin, *Social and Ethical Interpretations in Mental Development: A Study in Social Psychology*. New York, NY: Arno, 1973. (Original work published 1897.)
- [24] A. M. Turing, "Computing Machinery and Intelligence" *Mind*, vol. 59, no. 236, pp. 433-60, 1950.
- [25] Y. Akiwa, Y. Sugi, T. Ogata, and S. Sugano, "Imitation based human-robot interaction: -Roles of joint attention and motion prediction-," *Proceedings of the 14<sup>th</sup> International Workshop on Robot and Human Interactive Communication*, Kurashiki, Okayama, Japan, 2004, pp. 283-288.
- [26] A. Alissandrakis, C. L. Nehaniv, K. Dautenhahn, and J. Saunders, Evaluation of robot imitation attempts: Comparison of the system's and the human's perspectives, in *Proceedings of the 1<sup>st</sup> Annual Conference on Human-Robot Interaction*, Salt Lake City, UT, 2006, pp. 134-141.
- [27] C. Breazeal, and B. Scassellati, "Robots that imitate humans," *Trends in Cognitive Sciences*, vol. 6, no. 11, pp. 481-487, Nov. 2002.
- [28] D. Buchsbaum, B. Blumberg, C. Breazeal, and A. N. Meltzoff, "A simulation-theory inspired social learning system for interactive characters," in *Proceedings of the 14<sup>th</sup> International Workshop on Robot and Human Interactive Communication*, Nashville, TN, 2005, pp. 85-90.
- [29] K. Dautenhahn, and C. L. Nehaniv, Eds., *Imitation in Animals and Artifacts*. Cambridge, Mass.: The MIT Press, 2002.
- [30] D. Yamamoto, M. Doi, N. Matsuhira, H. Ueda, M. Kidode, "Behavior fusion in a robotic interface for practicality and familiarity: -Approach by simultaneous imitations-," *Proceedings of the 14<sup>th</sup> International Workshop on Robot and Human Interactive Communication*, Kuroshiki, Okayama, Japan, 2004, pp. 114-119.
- [31] P. H. Kahn, Jr., "Children's obligatory and discretionary moral judgments," *Child Development*, vol. 63, no. 2, pp. 416-430, Apr. 1992.
- [32] E. Turiel, *The Development of Social Knowledge*. Cambridge, England: Cambridge University Press, 1983.
- [33] E. Turiel, "Moral development," in W. Damon, Ed., *Handbook of Child Psychology*, 5th ed., vol. 3, N. Eisenberg (Ed.), *Social, Emotional, and Personality Development*. New York, NY: Wiley, 1998, pp. 863-932.
- [34] R. Dworkin, *Taking Rights Seriously*. Cambridge, MA: Harvard University Press, 1978.
- [35] Y. P. Mei, "Mo Tzu," in P. Edwards, Ed., *The Encyclopedia of Philosophy*, vol. 5. New York, NY: Macmillan, 1972, pp. 409-410.
- [36] P. H. Kahn, Jr., B. Friedman, D. R. Perez-Granados, and N. G. Freier, "Robotic pets in the lives of preschool children," *Interaction Studies*, in press.
- [37] G. F. Melson, P. H. Kahn, Jr., A. M. Beck, B. Friedman, T. Roberts, and E. Garrett, "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*, Portland, OR, USA, 2005, pp. 1649-1652.
- [38] R. L. Severson, and P. H. Kahn, Jr., "Social and moral judgments about pesticides and the natural environment: A developmental study with farm worker children," presented at the *Biennial Meeting of the Society for Research in Child Development*, Atlanta, GA, USA, April 7-10, 2005.
- [39] B. Friedman, and P. H. Kahn, Jr. "Human agency and responsible computing: Implications for computer system design," *Journal of Systems and Software*, vol. 17, no. 1, Jan. 1992, pp. 7-14.
- [40] B. Friedman, and L. Millett, "'It's the computer's fault' - Reasoning about computers as moral agents," in *Conference Companion of the Conference on Human Factors in Computing Systems*, Denver, CO, USA, 1995, pp. 226-227.
- [41] B. Friedman, and P. H. Kahn, Jr., "Human values, ethics, and design," in J. A. Jacko and A. Sears, Eds., *The Human-Computer Interaction Handbook*. Mahwah, NJ: Erlbaum, 2003, pp. 1177-1201.
- [42] B. Friedman, P. Lin, and J. K. Miller, "Informed consent by design," in L. Cranor and S. Garfinkel, Eds., *Designing Secure Systems that People Can Use*. Cambridge, MA: O'Reilly and Associates, 2006.
- [43] T. Fong, L. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, nos. 3-4, pp. 143-166, Mar. 2003.
- [44] Oxford English Dictionary. Oxford University Press, Oxford, England, 2004. Retrieved May 24, 2004 from [http://dictionary.oed.com/cgi/entry/00199182?single=1&query\\_type=word&queryword=reciprocal&edition=2e&first=1&max\\_to\\_show=10](http://dictionary.oed.com/cgi/entry/00199182?single=1&query_type=word&queryword=reciprocal&edition=2e&first=1&max_to_show=10)
- [45] D. J. Povinelli, *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works*. New York, NY: Oxford University Press, 2000.
- [46] M. Tomasello. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press, 2000.