Interaction Studies

Social Behaviour and Communication in Biological and Artificial Systems

VOLUME 8 ISSUE 3 2007

John Benjamins Publishing Company
Amsterdam / Philadelphia

SPECIAL ISSUE

Psychological Benchmarks of Human–Robot Interaction

Peter H. Kahn, Jr. and Karl F. MacDorman

Authenticity in the age of digital companions

Sherry Turkle
Massachusetts Institute of Technology

The first generation of children to grow up with electronic toys and games saw computers as their "nearest neighbors." They spoke of computers as rational machines and of people as emotional machines, a fragile formulation destined to be challenged. By the mid-1990s, computational creatures, including robots, were presenting themselves as "relational artifacts," beings with feelings and needs. One consequence of this development is a crisis in authenticity in many quarters. In an increasing number of situations, people behave as though they no longer privilege authentic emotion. This paper examines watershed moments in the history of human-machine interaction, focusing on the implications of relational artifacts for our collective perception of aliveness and for human-to-human relationships. For now, the exploration of human-robot encounters leads us to questions about the human purposes of believable digital companions that are evocative but not authentic.

Keywords: authenticity, Cog, Eliza, human–robot interaction, nurturance, Kismet, Paro, relational artifacts, self object, sociable robotics, Tamagotchi

With the advent of "thinking" machines, old philosophical questions about life and consciousness acquired new immediacy. Computationally rich software and, more recently, robots have challenged our values and caused us to ask new questions about ourselves (Turkle, 2005 [1984]). Are there some tasks, such as providing care and companionship, that only befit living creatures? Can a human being and a robot ever be said to perform the *same* task? In particular, how shall we assign value to what we have traditionally called relational authenticity? In their review of psychological benchmarks for human–robot interaction, Kahn et al. (2007) include authenticity as something robots can aspire to, but it is clear that from their perspective robots will be able to achieve it without sentience. Here, authenticity is situated on a more contested terrain.

Interaction Studies 8:3 (2007), 501–517.
ISSN 1572–0373 / E-ISSN 1572–0381

Eliza and the crisis of authenticity

Joseph Weizenbaum's computer program Eliza brought some of these issues to the fore in the 1960s. Eliza prefigured an important element of the contemporary robotics culture in that it was one of the first programs that presented itself as a *relational artifact*, a computational object explicitly designed to engage a user in a relationship (Turkle, 2001, 2004; Turkle, Breazeal, Dasté, & Scassellati, 2006; Turkle, Taggart, Kidd, & Dasté, 2006). Eliza was designed to mirror users' thoughts and thus seemed consistently supportive, much like a Rogerian psychotherapist. To the comment, "My mother is making me angry," Eliza might respond, "Tell me more about your family," or "Why do you feel so negatively about your mother?" Despite the simplicity of how the program works – by string matching and substitution – Eliza had a strong emotional effect on many who used it. Weizenbaum was surprised that his students were eager to chat with the program and some even wanted to be alone with it (Turkle, 2005 [1984]; Weizenbaum, 1976). What made Eliza a valued interlocutor? What matters were so private that they could only be discussed with a machine? Eliza not only revealed people's willingness to talk to computers but their reluctance to talk to other people. Students' trust in Eliza did not speak to what they thought Eliza would understand but to their lack of trust in the people who would understand.

This "Eliza effect" is apparent in many settings. People who feel that psychotherapists are silent or disrespectful may prefer to have computers in these roles (Turkle, 1995). "When you go to a psychoanalyst, well, you're already going to a robot," reports an MIT administrator. A graduate student confides that she would trade in her boyfriend for a "sophisticated Japanese robot," if the robot would produce "caring behavior." The graduate student says she relies on a "feeling of civility" in the house. If the robot could "provide the environment," she would be "happy to produce the illusion that there is somebody really with me." Relational artifacts have become evocative objects, objects that clarify our relationships to the world and ourselves (Turkle, 2005 [1984]; 2007). In recent years, they have made clear the degree to which people feel alone with each other. People's interest in them indicates that traditional notions of authenticity are in crisis.

Weizenbaum came to see students' relationships with Eliza as immoral, because he considered human understanding essential to the confidences a patient shares with a psychotherapist. Eliza could not understand the stories it was being told; it did not care about the human beings who confided in it. Weizenbaum found it disturbing that the program was being treated as more than a parlor game. If the software *elicited* trust, it was only by tricking those who used it. From this viewpoint, if Eliza was a benchmark, it was because the software marked a crisis in

authenticity: people did not care if their life narratives were *really* understood. The act of telling them created enough meaning on its own.

When Weizenbaum's book that included his highly charged discussion of reactions to Eliza was published in 1976, I was teaching courses with him at MIT on computers and society. At that time, the simplicity and transparency of how the program worked helped Eliza's users recognize the chasm between program and person. The gap was clear as was how students bridged it with attribution and desire. They thought, "I will talk to this program *as if* it were a person." Hence, Eliza seemed to me no more threatening than an interactive diary. But I may have underestimated the quality of the connection between person and machine. To put it too simply, when a machine shows interest in us, it pushes our "Darwinian buttons" (Turkle, 2004) that signal it to be an entity appropriate for relational purposes. The students may not have been pretending that they were chatting with a person. They may just have been happy to talk to a machine.

This possibility is supported by new generations of digital creatures that create a greater sense of mutual relating than Eliza, but *have no greater understanding of the situation of the human being in the relationship*. The relational artifacts of the past decade, *specifically designed to make people feel understood*, provide more sophisticated interfaces, but they are still without understanding.

Some of these relational artifacts are very simple in what they present to the user, such as the 1997 Tamagotchi, a virtual creature that inhabits a tiny LCD display. Some of them are far more complex, such as Kismet, developed at the MIT Artificial Intelligence Laboratory, a robot that responds to facial expressions, vocalizations, and tone of voice. From 1997 to the present I have conducted field research with these relational artifacts and also with Furbies, Aibos, My Real Babies, Paros, and Cog. What these machines have in common is that they display behaviors that make people feel as though they are dealing with sentient creatures that care about their presence. These Darwinian buttons, these triggering behaviors, include making eye contact, tracking an individual's movement in a room, and gesturing benignly in acknowledgment of human presence. People who meet these objects feel a desire to nurture them. And with this desire comes the fantasy of reciprocation. People begin to care for these objects and want the objects to care about them.

In the 1960s and 1970s, confiding in Eliza meant ignoring the program's mechanism so that it seemed mind-like and thus worthy of conversation. Today's interfaces are designed to make it easier to ignore the mechanical aspects of the robots and think of them as nascent minds.

In a 2001 study, my colleagues and I tried to make it harder for a panel of thirty children to ignore machine mechanism when relating to the Cog robot at the MIT AI Lab (Turkle, Breazeal, Dasté, & Scassellati, 2006). When first presented with the

robot, the children (from age 5 to 13) delighted in its presence. They treated it as a creature with needs, interests, and a sense of humor. During the study, one of Cog's arms happened to be broken. The children were concerned, tried to make Cog more comfortable, wanted to sing and dance to cheer it up, and in general, were consistently solicitous of its "wounds." Then, for each child, there was a session in which Cog was demystified. Each child was shown Cog's inner workings, revealing the robot as "mere mechanism." During these sessions, Brian Scassellati, Cog's principal developer, painstakingly explained how Cog could track eye movement, follow human motion, and imitate behavior. In the course of a half hour, Cog was shown to be a long list of instructions scrolling on a computer screen. Yet, within minutes of this demonstration, children were back to relating to Cog as a creature and playmate, vying for its attention. Similarly, when we see the functional magnetic resonance imaging (fMRI) of a person's brain, we are not inhibited in our ability to relate to that person as a meaning-filled other. The children, who so hoped for Cog's affection, are being led by the human habit of making assumptions based on perceptions of behavior. But the robot in which the children were so invested did not care about them. As was the case for Eliza, desire bridged the distance between the reality of the program and the children's experience of it as a sentient being. Kahn et al. (2007) might classify this bridging as a "psychological benchmark," but to return to the Eliza standard, if it is a benchmark, it is only in the eye of the beholder. To have a relationship, the issue is not only what the human feels but what the robot feels.

Human beings evolved in an environment that did not require them to distinguish between authentic and simulated relationships. Only since the advent of computers have people needed to develop criteria for what we consider to be "authentic" relationships, and for many people the very idea of developing these criteria does not seem essential. For some, the idea of computer companionship seems natural; for others, it is close to obscene. Each group feels its position is self-evident. Philosophical assumptions become embedded in technology; radically different views about the significance of authenticity are at stake. As robots become a part of everyday life, it is important that these differences are clearly articulated and discussed.

At this point, it seems helpful to reformulate a notion of benchmarks that puts authenticity at center stage. In the presence of relational artifacts and, most recently, robotic creatures, people are having feelings that are reminiscent of what we would call trust, caring, empathy, nurturance, and even love, *if* they were being called forth by encounters with people. But it seems odd to use these words to describe benchmarks in human-robot encounters, because we have traditionally reserved them for relationships in which all parties were capable of feeling them – that is, where all parties were people. With robots, people are acting out "both halves" of complex relationships, projecting the robot's side as well as their own. Of course, we can also behave this way when interacting with people who refuse to engage with us, but people are at least *capable* of reciprocation. We can be disappointed in people, but at least we are disappointed about genuine potential. For robots, the issue is not disappointment, because the idea of reciprocation is pure fantasy.

It belongs to the future to determine whether robots could ultimately "deserve" the emotional responses they are now eliciting. For now, the exploration of human-robot encounters leads us instead to questions about the human purposes of digital companions that are evocative but not relationally authentic.

The recent history of computation and its psychological benchmarks

We already know that the "intimate machines" of the computer culture have shifted how children talk about what is and is not alive (Turkle, 2005 [1984], 1995; Turkle, Breazeal, Dasté, & Scassellati, 2006; Kahn, Friedman, Pérez-Granados, & Freier, 2006). As a psychological benchmark, *aliveness* has presented a moving target. For example, children use different categories to talk about the aliveness of "traditional" objects versus computational games and toys. A traditional wind-up toy was considered "not alive" when children realized that it did not move of its own accord (Piaget, 1960). The criterion for aliveness, autonomous motion, was operationalized in the domain of physics.

In the late 1970s and early 1980s, faced with computational media, there was a shift in how children talked about aliveness. Their language became psychological. By the mid-1980s, children classified computational objects as alive if the objects could *think* on their own. Faced with a computer toy that could play tic-tac-toe, children's determination of aliveness was based on the object's *psychological rather than physical autonomy*. As children attributed psychological autonomy to computational objects, they also split consciousness and life (Turkle, 2005[1984]). This enabled children to grant that computers and robots might have consciousness (and thus be aware both of themselves and of us) without being alive.

This first generation of children who grew up with computational toys and games classified them as "sort of alive," in contrast to the other objects of the playroom (Turkle, 2005 [1984]). Beyond this, they came to classify computational objects as people's "nearest neighbors" because of the objects' intelligence. People were different from these neighbors because of their emotions. Thus, children's formulation was that computers were "intelligent machines," distinguished from people who had capacities as "emotional machines." I anticipated that later generations of children would find other formulations as they learned more about

computers. They might, for example, see through the apparent "intelligence" of the machines by developing a greater understanding of how they were created and operated. As a result, children might be less inclined to give computers philosophical importance. However, in only a few years, both children and adults would quickly learn to overlook the internal workings of computational objects and forge relationships with them based on their behaviour (Turkle, 1995, 2005 [1984]).

The lack of interest in the inner workings of computational objects was reinforced by the appearance in mainstream American culture of robotic creatures that presented themselves as having both feelings and needs. By the mid-1990s, people were not alone as "emotional machines." This new generation of objects was designed to approach the boundaries of humanity not so much with its "smarts" as with its sociability (Kiesler & Sproull, 1997; Parise, Kiesler, Sproull, & Waters, 1999; Reeves & Nass, 1999).

The first relational artifacts to enter the American marketplace were virtual creatures known as Tamagotchis that lived on a tiny LCD screen housed in a small plastic egg. The Tamagotchis — a toy fad of the 1997 holiday season — were presented as creatures from another planet that needed human nurturance, both physical and emotional. An individual Tamagotchi would grow from child to healthy adult if it was cleaned when dirty, nursed when sick, amused when bored, and fed when hungry. A Tamagotchi, while it lived, needed constant care. If its needs were not met, it would expire. Children became responsible parents; they enjoyed watching their Tamagotchis thrive and did not want them to die. During school hours, parents were enlisted to care for the Tamagotchis; beeping Tamagotchis became background noise during business meetings. Although primitive as relational artifacts, the Tamagotchis demonstrated a fundamental truth of a new human–machine psychology. When it comes to bonding with computers, *nurturance* is the "killer app" (an application that can eliminate its competitors). When a digital creature entrains people to play parent, they become attached. They feel connection and even empathy.

It is important to distinguish feelings for relational artifacts from those that children have always had for the teddy bears, rag dolls, and other inanimate objects they turn into imaginary friends. According to the psychoanalyst D.W. Winnicott, objects such as teddy bears mediate between the infant's earliest bonds with the mother, who is experienced as inseparable from the self, and other people, who will be experienced as separate beings (Winnicott, 1971). These objects are known as "transitional," and the infant comes to know them as both almost-inseparable parts of the self and as the first "not me" possessions. As the child grows, these transitional objects are left behind, but the effects of early encounters with them are manifest in the highly charged intermediate space between the self and certain

objects in later life, objects that become associated with religion, spirituality, the perception of beauty, sexual intimacy, and the sense of connection with nature.

How are today's relational artifacts different from Winnicott's transitional objects? In the past, the power of early objects to play a transitional role was tied to how they enabled a child to project meanings onto them. The doll or teddy bear presents an unchanging and passive presence. Today's relational artifacts are decidedly more active. With them, children's expectations that their dolls want to be hugged, dressed, or lulled to sleep come not from children's projections of fantasy onto inert playthings, but from such things as a digital doll or robot's inconsolable crying or exclamation "Hug me!" or "It's time for me to get dressed for school!" So when relational artifacts prospered under children's care in the late 1990s and early 2000s, children's discourse about the objects' aliveness subtly shifted. Children came to describe relational artifacts in the culture (first Tamagotchis, then Furbies, Aibos, and My Real Babies) as alive or "sort of alive," not because of what these objects could do (physically *or* cognitively) but because of the children's emotional connection to the objects and their fantasies about how the objects might be feeling about them. The focus of the discussion about whether these objects might be alive moved from the psychology of projection to the psychology of engagement, from Rorschach (i.e., projection, as on an inkblot) to relationship, from creature competency to creature connection.

In the early 1980s, I met 13-year-old Deborah, who described the pleasures of projection onto a computational object as putting "a piece of your mind into the computer's mind and coming to see yourself differently" (2005 [1984]). Twenty years later, 11-year-old Fara reacts to a play session with Cog, the humanoid robot at MIT, by saying that she could never get tired of the robot, because "it's not like a toy because you can't teach a toy; it's like something that's part of you, you know, something you love, kind of like another person, like a baby" (Turkle, Breazeal, Dasté, & Scassellati, 2006). The contrast between these two responses reveals a shift from projection onto an object to engagement with a subject.

Engagement with a subject

In the 1980s, debates in artificial intelligence centered on whether machines could be intelligent. These debates were about the objects themselves, what they could and could not do and what they could and could not be (Searle, 1980; Dreyfus, 1986; Winograd, 1986). The questions raised by relational artifacts are not so much about the machines' capabilities but our vulnerabilities — not about whether the objects *really* have emotion or intelligence but about what they evoke in us. For when we are asked to care for an object, when the cared-for object thrives and

offers us its "attention" and "concern," we not only experience it as intelligent, but more importantly, we feel a heightened connection to it.

Even very simple relational artifacts can provoke strong feelings. In one study of 30 elementary school age children who were given Furbies to take home (Turkle, 2004), most had bonded emotionally with their Furby and were convinced that they had taught the creature to speak English. (Each Furby arrives "speaking" only Furbish, the language of its "home planet" and over time "learns" to speak English.) Children became so attached to their particular Furby that when the robots began to break, most refused to accept a replacement. Rather, they wanted their own Furby "cured." The Furbies had given the children the feeling of being successful caretakers, successful parents, and they were not about to "turn in" their sick babies.

The children had also developed a way of talking about their robots' "aliveness" that revealed how invested the children had become in the robots' well being. There was a significant integration of the discourses of aliveness and attachment. Ron, six, asks, "Is the Furby alive? Well, something this smart should have arms... it might want to pick up something or to hug me." When Katherine, five, considers Furby's aliveness, she, too, speaks of her love for her Furby and her confidence that it loves her back: "It likes to sleep with me." Jen, nine, admits how much she likes to take care of her Furby, how comforting it is to talk to it (Turkle, 2004). These children are learning to have expectations of emotional attachments to robots in the same way that we have expectations about our emotional attachments to people. In the process, the very meaning of the word *emotional* is changing. Children talk about an "animal kind of alive and a Furby kind of alive." Will they also talk about a "people kind of love" and a "robot kind of love?"

In another study, 60 children from age 5 to 13 were introduced to Kismet and Cog (Turkle, Breazeal, Dasté, & Scassellati, 2006). During these first encounters, children hastened to put themselves in the role of the robots' teachers, delighting in any movement (for Cog), vocalization or facial expression (for Kismet) as a sign of robot approval. When the robots showed imitative behavior they were rewarded with huggs and kisses. One child made clay treats for Kismet. Another told Kismet, "I'm going to take care of you and protect you against all evil." Another decided to teach the robots sign language, because they clearly had trouble with spoken English, and to begin with the signs for "house," "eat," and "I love you."

In a study of robots and the elderly in Massachusetts nursing homes, emotions ran similarly high (Turkle, Taggart, et al., 2006). Jonathan, 74, responds to My Real Baby, a robot baby doll he keeps in his room, by wishing it were a bit smarter, because he would prefer to talk to a robot about his problems than to a person. "The robot wouldn't criticize me," he says. Andy, also 74, says that his My Real Baby, which responds to caretaking by developing different states of "mind," resembles

his ex-wife Rose: "something in the eyes." He likes chatting with the robot about events of the day. "When I wake up in the morning and see her face [the robot's] over there, it makes me feel so nice, like somebody is watching over me."

In Philip K. Dick's (1968) classic story, *Do Androids Dream of Electric Sheep* (a novel that most people know through its film adaptation *Blade Runner*), androids act like people, developing emotional connections with each other and the desire to connect with humans. *Blade Runner*'s hero, Deckard, makes his living by distinguishing machines from human beings based on their reactions to a version of the Turing Test for distinguishing computers from people, the fictional Voight-Kampff test. What is the difference, asks the film, between a real human and an almost-identical object? Deckard, as the film progresses, falls in love with the near-perfect simulation, the android Rachael. Memories of a human childhood and the knowledge that her death is certain make her seem deeply human. By the end of the film, we are left to wonder whether Deckard himself may also be an android who is unaware of his status. Unable to resolve this question, viewers are left cheering for Deckard and Rachael as they escape to whatever time they have remaining, in other words, to the human condition. The film leaves us to wonder whether, by the time we face the reality of computational devices that are indistinguishable from people, and thus able to pass our own Turing test, we will no longer care about the test. By then, people will love their machines and be more concerned about their machines' happiness than their test scores.

This conviction is the theme of a short story by Brian Aldiss (2001), "Supertoys Last All Summer Long," that was made into the Steven Spielberg film *AI: Artificial Intelligence*. In *AI*, scientists build a humanoid robot, David, that is programmed to love. David expresses his love to Monica, the woman who has adopted him. Our current experience with relational artifacts suggests that the pressing issue raised by the film is not the potential reality of a robot that "loves," but the feelings of the adoptive mother, whose response to the machine that asks for nurturance is a complex mixture of attachment and confusion. Cynthia Breazeal's experience at the MIT AI Lab offers an example of how such relationships might play out in the near term. Breazeal led the design team for Kismet, the robotic head designed to interact with people as a two-year-old might. She was Kismet's chief programmer, tutor, and companion. Breazeal developed what might be called a maternal connection with Kismet; when she graduated from MIT and left the AI Lab where she had completed her doctoral research, the tradition of academic property rights demanded that Kismet remain in the laboratory that had paid for its development. Breazeal described a sharp sense of loss. Building a new Kismet would not be the same.

Breazeal worked with me on the "first encounters" study of children interacting with Kismet and Cog during the summer of 2001, the last time she would have access to Kismet. It is not surprising that separation from Kismet was not easy for

Breazeal, but more striking was how hard it was for those around Kismet to imagine the robot without her. One 10-year-old who overheard a conversation among graduate students about how Kismet would remain behind objected, "But Cynthia is Kismet's mother."

It would be facile to compare Breazeal's situation to that of Monica, the mother in Spielberg's *AI*, but Breazeal is, in fact, one of the first adults to have the key human experience portrayed in that film, sadness caused by separation from a robot to which one has formed an attachment based on nurturance. What is at issue is the emotional effect of Breazeal's experience as a "caregiver." In a very limited sense, Breazeal "brought up" Kismet. But even this very limited experience provoked strong emotions. Being asked to nurture a machine *constructs us* as its parents. Although the machine may only have simulated emotion, the feelings it evokes are real. Successive generations of robots may well be enhanced with the specific goal of engaging people in affective relationships by asking for their nurturance. The feelings they elicit will reflect human vulnerabilities more than machine capabilities (Turkle, 2003).

Imitation beguiles

In the case of the Eliza program, imitation beguiled users. Eliza's ability to mirror and manipulate what it was told was compelling, even if primitive. Today, designers of relational artifacts are putting this lesson into practice by developing robots that appear to empathize with people by mimicking their behavior, mirroring their moods (Shibata, 2004). But again, as one of Kahn et al.'s (2007) proposed benchmarks, imitation is less psychologically important as a measure of machine ability than of human susceptibility to this design strategy.

Psychoanalytic self psychology helps us think about the human effects of this kind of mimicry. Heinz Kohut describes how some people may shore up their fragile sense of self by turning another person into a "self object" (Ornstein, 1978). In this role, the other is experienced as part of the self, and as such must be attuned to the fragile individual's inner state. Disappointments inevitably follow. Someday, if relational artifacts can give the impression of aliveness and not disappoint, they may have a "comparative advantage" over people as self objects and open up new possibilities for narcissistic experience. For some, predictable relational artifacts are a welcome substitute for the always-resistant human material. What are the implications of such substitutions? Do we want to shore up people's narcissistic possibilities?

Over 25 years ago, the Japanese government projected that there would not be enough young people to take care of their older population. They decided that instead of having foreigners take care of their elderly, they would build robots. Now, some of these robots are being aggressively marketed in Japan, some are in development, and some are poised for introduction in American settings.

US studies of the Japanese relational robot Paro have shown that in an elder-care setting, administrators, nurses, and aides are sympathetic toward having the robot around (Turkle, Taggart, et al., 2006). It gives the seniors something to talk about as well as something new to talk to. Paro is a seal-like creature, advertised as the first "therapeutic robot" for its apparently positive effects on the ill, the elderly, and the emotionally troubled (Shibata, 2004). The robot is sensitive to touch, can make eye contact by sensing the direction of a voice, and has states of "mind" that are affected by how it is treated. For example, it can sense if it is being stroked gently or aggressively. The families of seniors also respond warmly to the robot. It is not surprising that many find it easier to leave elderly parents playing with a robot than staring at a wall or television set.

In a nursing home study on robots and the elderly, Ruth, 72, is comforted by the robot Paro after her son has broken off contact with her (Turkle, Taggart, et al., 2006). Ruth, depressed about her son's abandonment, comes to regard the robot as being equally depressed. She turns to Paro, strokes him, and says, "Yes, you're sad, aren't you. It's tough out there. Yes, it's hard." Ruth strokes the robot once again, attempting to comfort it, and in so doing, comforts herself.

This transaction brings us back to many of the questions about authenticity posed by Eliza. If a person *feels* understood by an object lacking sentience, whether that object be an imitative computer program or a robot that makes eye contact and responds to touch, can that illusion of understanding be therapeutic? What is the status — therapeutic, moral, and relational — of the simulation of understanding? If a person claims they feel better after interacting with Paro, or prefers interacting with Paro to interacting with a person, what are we to make of this claim? It seems rather a misnomer to call this a "benchmark in interaction." If we use that phrase we must discipline ourselves to keep in mind that Paro understands nothing, senses nothing, and cares nothing for the person who is interacting with it. The ability of relational artifacts to inspire "the feeling of relationship" is not based on their intelligence, consciousness, or reciprocal pleasure in relating, but on their ability to push our Darwinian buttons, by making eye contact, for example, which causes people to respond *as if* they were in a relationship.

If one carefully restricts Kahn et al.'s (2007) benchmarks to refer to feelings elicited in people, it is possible that such benchmarks as *imitation, mutual relating,* and *empathy* might be operationalized in terms of machine actions that could be coded and measured. In fact, the work reviewed in this paper suggests the addition of the attribution of *aliveness, trust, caring, empathy, nurturance,* and *love* to a list of benchmarks, because people are capable of feeling all these things for a robot

and believing a robot feels them in return. But these benchmarks are very different from psychological benchmarks that measure authentic experiences of relationship. What they measure is the human perception of what the machine *would be experiencing* if a person (or perhaps an animal) evidenced the behaviors shown by the machine.

Such carefully chosen language is reminiscent of early definitions of AI. One famous formulation proposed by Marvin Minsky had it that "artificial intelligence is the science of making machines do things that would require intelligence if done by [people]" (Minsky, 1968, p. v). There is a similar point to be made in relation to Kahn et al.'s (2007) benchmarks. To argue for a benchmark such as Buber's (1970) "I-You" relating, or even to think of adding things such as empathy, trust, caring, and love to a benchmark list, is either to speak *only* in terms of human attribution or to say, "The robot is exhibiting behavior that would be considered caring if performed by a person (or perhaps an animal)."

Over the past 50 years, we have built not only computers but a computer culture. In this culture, language, humor, art, film, literature, toys, games, and television have all played their role. In this culture, the subtlety of Minsky's careful definition of AI dropped out of people's way of talking. With time, it became commonplace to speak of the products of AI as though they had an inner life and inner sense of purpose. As a culture, we seem to have increasingly less concern about how computers operate internally. Ironically, we now term things "transparent" if we know how to make them work rather than if we know how they work. This is an inversion of the traditional meaning of the word transparency, which used to mean something like being able to "open the hood and look inside." People take interactive computing, including interactive robots, "at interface value" (Turkle, 1995, 2005 [1984]). These days, we are not only building robots, but a robot culture. If history is our guide, we risk coming to speak of robots as though they also have an inner life and inner sense of purpose. We risk taking our benchmarks at face value.

In the early days of artificial intelligence, people were much more protective of what they considered to be exclusively human characteristics, expressing feelings that could be characterized in the phrase: "Simulated thinking is thinking, but simulated feeling is not feeling, and simulated love is never love" (Turkle, 2005 [1984]). People accepted the early ambitions of artificial intelligence, but drew a line in the sand. Machines could be cognitive, but no more. Nowadays, we live in a computer culture where there is regular talk of affective computing, sociable machines, and flesh and machine hybrids (Picard, 1997; Breazeal, 2002; Brooks, 2002). Kahn et al.'s (2007) benchmarks reflect this culture. There has been an erosion of the line in the sand, both in academic life and in the wider culture.

What may provoke a new demarcation of where computers should not go are robots that make people uncomfortable, robots that come *too* close to the human. As robotics researchers create humanlike androids that strike people as uncanny, they strike people as somehow "not right" (MacDorman & Ishiguro, 2006a, 2006b). Current analyses of uncanny robot interactions are concerned with such things as appearance, motion quality, and interactivity. But as android work develops, it may be questions of values and authenticity that turn out to be at the heart of human concerns about these new objects.

Freud wrote of the uncanny as the long familiar seeming strangely unfamiliar, or put another way, the strangely unfamiliar embodying aspects of the long familiar (Freud, 1960 [1919]). In every culture, confrontation with the uncanny provokes new reflection. Relational artifacts are the new uncanny in our computer culture. If our experience with relational artifacts is based on the fiction that they know and care about us, can the attachments that follow be good for us? Or might they be good for us in the "feel good" sense, but bad for us as moral beings? The answers to such questions do not depend on what robots can do today or in the future. These questions ask what *we* will be like, what kind of people *we* are becoming as we develop increasingly intimate relationships with machines.

The purposes of living things

Consider this moment: Over the school break of Thanksgiving 2005, I take my 14-year-old daughter to the Darwin exhibit at the American Museum of Natural History in New York. The exhibit documents Darwin's life and thought and presents the theory of evolution as the central truth that underpins contemporary biology. At the entrance to the exhibit lies a Galapagos turtle, a seminal object in the development of evolutionary theory. The turtle rests in its cage, utterly still. "They could have used a robot," comments my daughter. Utterly unconcerned with the animal's authenticity, she thinks it a shame to bring the turtle all this way to put it in a cage for a performance that draws so little on its "aliveness."

In talking with other parents and children at the exhibit, my question, "Do you care that the turtle is alive?" provokes a variety of responses. A 10-year-old girl would prefer a robot turtle, because aliveness comes with aesthetic inconvenience: "Its water looks dirty, gross." More often, the museum's visitors echoed my daughter's sentiment that, in this particular situation, actual aliveness is unnecessary. A 12-year-old girl opines, "For what the turtles do, you didn't have to have the live ones." The girl's father is quite upset: "But the point is that they are real. That's the whole point." "If you put in a robot instead of the live turtle, do you think people should be told that the turtle is not alive?" I ask. "Not really," say several children.

Apparently, data on "aliveness" can be shared on a "need to know" basis, for a purpose. But what are the purposes of living things? These children struggle to find any. They are products of a culture in which human contact is routinely replaced by virtual life, computer games, and now relational artifacts.

The Darwin exhibit emphasizes authenticity; on display is the actual magnifying glass that Darwin used, the actual notebooks in which he recorded his observations, and the very notebook in which he wrote the famous sentences that first described his theory of evolution. But, ironically, in the children's reactions to the inert but alive Galapagos turtle, the idea of the "original" is in crisis.

Sorting out our relationships with robots brings us back to the kinds of challenges that Darwin posed to his generation regarding human uniqueness. How will interacting with relational artifacts affect how people think about what, if anything, makes people special? Ancient cultural axioms that govern our concepts about aliveness and emotion are at stake. Robots have already shown the ability to give people the illusion of relationship: Paro convinced an elderly woman that it empathized with her emotional pain; students ignored the fact that Eliza was a parrot-like computer program, choosing instead to accept its artificial concern. Meanwhile, examples of children and the elderly exchanging tenderness with robotic pets bring science fiction and techno-philosophy into everyday life.

Ultimately, the question is not whether children will love their robotic pets more than their animal pets, but rather, what loving will come to mean. Going back to the young woman who was ready to turn in her boyfriend for a "sophisticated Japanese robot," is there a chance that human relationships will just seem too *hard?* There may be some who would argue that the definition of relationships should broaden to accommodate the pleasures afforded by cyber-companionship, however inauthentic. Indeed, people's positive reaction to relational artifacts would suggest that the term is being contested. In the culture of simulation, authenticity is for us what sex was to the Victorians: taboo and fascination, threat and preoccupation.

Perhaps in the distant future, the difference between human beings and robots will seem purely philosophical. A simulation of the quality of Rachael in *Blade Runner* could inspire love on a par with what we feel toward people. In thinking about the meaning of love, however, we need to know not only what the people are feeling but what the robots are feeling. We are easily seduced; we easily forget what they are; we easily forget what we have made.

As I was writing this paper, I discussed it with a former colleague, Richard, who had been left severely disabled by an automobile accident. He is now confined to a wheelchair in his home and needs nearly full-time nursing help. Richard was interested in robots being developed to provide practical help and companionship to people in his situation. His reaction to the idea was complex. He began by saying, "Show me a person in my shoes who is looking for a robot, and I'll show you someone who is looking for a person and can't find one," but then he made the best possible case for robotic helpers. He turned the conversation to human cruelty: "Some of the aides and nurses at the rehab center hurt you because they are unskilled and some hurt you because they mean to. I had both. One of them, she pulled me by the hair. One dragged me by my tubes. A robot would never do that," he said. "But you know in the end, that person who dragged me by my tubes had a story. I could find out about it."

For Richard, being with a person, even an unpleasant, sadistic person, made him feel that he was still alive. It signified that his way of being in the world still had a certain dignity, for him the same as authenticity, even if the scope and scale of his activities were radically reduced. This helped sustain him. Although he would not have wanted his life endangered, he preferred the sadist to the robot. Richard's perspective on living is a cautionary word to those who would speak too quickly or simply of purely technical benchmarks for our interactions. What is the value of interactions that contain no understanding of us and that contribute nothing to a shared store of human meaning? These are not questions with easy answers, but questions worth asking and returning to.

Acknowledgments

Research reported in this paper was funded by an NSF ITR grant "Relational Artifacts" (Turkle 2001) award number SES-0115668, by a grant from the Mitchell Kapor Foundation, and by a grant from the Intel Corporation.

References

Aldiss, B. W. (2001). *Supertoys last all summer long and other stories of future time.* New York: St. Martin.

Breazeal, C. (2002). *Designing sociable robots.* Cambridge, MA: MIT Press.

Brooks, R. A. (2002). *Flesh and machines: How robots will change us.* New York: Pantheon Books.

Buber, M. (1970). *I and thou.* New York: Touchstone.

Dick, P. K. (1968). Do androids dream of electric sheep? Garden City, NY: Doubleday.

Dreyfus, H. L. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York: Free Press.

Freud, S. (1960 [1919]). The uncanny. In J. Strachey (Transl., Ed.), *The standard edition of the complete psychological works of Sigmund Freud* (vol. 17, pp. 219–252). London: The Hogarth Press.

Kahn, P. H., Jr., Friedman, B., Pérez-Granados, D. R., & Freier, N. G. (2006). Robotic pets in the lives of preschool children. *Interaction Studies, 7*(3), 405–436.

Kahn, P. H., Jr., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., & Miller, J. (2007). What is a human? — Toward psychological benchmarks in the field of human–robot interaction. *Interaction Studies 8:3*. (This issue)

Kiesler, S. & Sproull, L. (1997). Social responses to "social" computers. In B. Friedman (Ed.), *Human values and the design of technology*. Stanford, CA: CLSI Publications.

MacDorman, K. F. & Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies, 7*(3), 297–337.

MacDorman, K. F. & Ishiguro, H. (2006). Opening Pandora's uncanny box: Reply to commentaries on "The uncanny advantage of using androids in social and cognitive science research." *Interaction Studies, 7*(3), 361–368.

Ornstein, P. H. (Ed). (1978). *The search for the self: Selected writings of Heinz Kohut (1950–1978)* (vol. 2). New York: International Universities Press.

Parise, S., Kiesler, S., Sproull, L., & Waters, K. (1999). Cooperating with life-like interface agents. *Computers in Human Behavior, 15*(2), 123–142.

Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Piaget, J. (1960 [1929]). *The child's conception of the world* (transl. J. & A. Tomlinson). Totowa, N.J.: Littlefield, Adams.

Reeves, B. & Nass, C. (1999). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge: Cambridge University Press.

Searle, J. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3, 417–424.

Shibata, T. (2004). An overview of human interactive robots for psychological enrichment. *Proceedings of the IEEE, 92*(11), 1749–1758.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. New York: Simon and Schuster.

Turkle, S. (2001). *Relational artifacts*. Proposal to the National Science Foundation SES-01115668.

Turkle, S. (2003). Technology and human vulnerability. *The Harvard Business Review*, September.

Turkle, S. (2004). Whither Psychoanalysis in the Computer Culture? *Psychoanalytic Psychology, 21*(1), 16–30.

Turkle, S. (2005 [1984]). *The second self: Computers and the human spirit*. Cambridge, MA: MIT Press.

Turkle, S. (2006). Diary. *The London Review of Books, 8*(8), April 20.

Turkle, S., Breazeal, C., Dasté, O., & Scassellati, B. (2006). First encounters with Kismet and Cog: Children's relationship with humanoid robots. In P. Messaris & L. Humphreys (Eds.), *Digital media: Transfer in human communication*. New York: Peter Lang.

Turkle, S., Taggart, W., Kidd, C. D. & Dasté, O. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. *Connection Science, 18*(4), 347–361.

Turkle, S. (Ed). (2007) *Evocative objects: Things we think with*. Cambridge, MA: MIT Press.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. San Francisco, CA: W. H. Freeman.

Winnicott, D. W. (1971). *Playing and reality*. New York: Basic Books.

Winograd, T. & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Norwood, NJ: Ablex.

Author's address

Sherry Turkle
Massachusetts Institute of Technology
Program in Science, Technology, and Society
E51-296C
Cambridge, MA 02139 USA

sturkle@media.mit.edu

About the author

Sherry Turkle is the Abby Rockefeller Mauzé Professor of the Social Studies of Science and Technology at the Massachusetts Institute of Technology and Director of its Initiative on Technology and Self. She received a Ph.D. in Sociology and Personality Psychology from Harvard University. She is a licensed clinical psychologist and a member of the Boston Psychoanalytic Society. Her research interests include the subjective side of the computer presence and the social and psychological impact of relational artifacts.