19 Human-Robot Interaction

Tony Belpaeme

19.1 Introduction

Human-robot interaction (HRI) studies the interaction between people and robotic systems. While robots are traditionally operated using user interfaces gleaned from human-computer interaction, such as control panels or screen-based interfaces, there is potential to move toward more natural modes of interaction. These will, to a large extent, be modeled on how people interact with each other and are composed of verbal and nonverbal ways of interacting.

HRI is a broad church: at one end of the spectrum, it studies how an operator can control one or more robotic systems through traditional methods and sometimes focuses on the cognitive load imposed by controlling one or more robots. For example, if an operator coordinates a handful of semiautonomous drones during a search and rescue operation, how can the cognitive load on the operator be optimized to maximize the efficiency of the overall mission (e.g., Goodrich et al. 2011)? On the other end of the spectrum of HRI, one finds research into natural interaction between humans and robots. This field is also known as *social robotics*, and the large majority of research efforts in HRI concentrate on it (Bartneck et al. 2020). The holy grail of social HRI, of course, is the natural and intuitive interaction between people and artificial systems. On one hand, this is a technical effort, with results in social signal processing, artificial intelligence, and robotics coming together to create social robots. But social robotics offers a unique opportunity to study how people respond and interact with artificial social agents. Social robots take up a singular position in agents we interact with. The interaction between people has, of course, been the subject of extensive study for more than a century, and the interaction between animals and people has been researched at length, but robots are a new and, until recently, unexplored "species." Until recently, we have known very little about how people interact with robots, and our relation and interaction with robots is continuously evolving. Culture, media, education, context, and exposure change our attitudes toward robots and the ways in which we interact with them. When we meet a robot, several automatic social responses kick in that color our interaction with the robot; these responses evolved or developed to interact with other humans and often transfer to our interaction with robots.

This is not unique to robots. We treat all technology to some extent as if it is humanlike, something known as anthropomorphization, which Clifford Nass called the "media equation." We relate to media—computers, printers, mobile phones, and of course robots—as if they are human (Reeves and Nass 1996). Everyone has at one time or another muttered at their computer when it crashed or cursed their printer when the paper jammed, but the media equation theory takes things a little further by claiming that we not only respond to these media as if they were persons but ascribe personal qualities to each, such as a personality, expertise, and even gender. And we often do so without being aware of it. The media equation is taken to the extreme in social robots, as the appearance of the robot and its behavior (the things it does) have been carefully designed to elicit a strong social response from us.

19.2 Cognitive and Neuroscientific Insights Informing HRI

Social psychology is immediately relevant to the design of social robots, and knowingly or not, designers and programmers of social robots take concepts and theories from social psychology into consideration when building robots. Failing to do so usually results in a disappointing HRI. Whether you wish to create a friendly robot or a horror experience, you will rely on fundamentals from social psychology when designing the appearance of your robot and its interaction.

The media equation predicts that people will perceive and treat robots in a humanlike way, but the fact that we readily interpret animated objects as having humanlike emotions and intentions has been known for a long time. Fritz Heider and Marianne Simmel (1944), two psychologists working together in the United States, published an influential paper titled "An Experimental Study of Apparent Behavior" in which they described a simple and elegant experiment: They asked people to describe short film clips of moving geometric figures, such as circles and triangles. The figures were animated by hand and seemed to play out a short story. Everyone who saw the videos ascribed emotions and intentions to the figures. The original videos from the 1940s can still be found online, and even now when seeing the videos, people readily see the figures having emotions, intentions, and motivations, and they see a narrative unfold over the few minutes of video runtime. This is our social brain interpreting the world around it and, specifically, our theory of mind—our ability to attribute mental states to others and ourself—overinterpreting moving geometric figures. This concept has been gratefully used by animators, and some striking examples exist of very minimalist animation films that show that very little is needed to nudge our social brain into interpreting simple shapes and movement as having agency (Thomas and Johnston 1995). If you have ever observed a vacuuming robot moving around the room, you have probably been struck by its animallike appearance as it scuttles around the room, gently bumping into furniture and working hard at getting specks of dirt from the floor. These robots are not designed to be social, and yet they still evoke a strong social response in us. In social robots, designers add elements such as a head, eyes, and reactive responses to evoke a strong social response in people.

One such social response on which designers rely is *pareidolia*: the tendency to see human or animal forms in objects, such as dogs in clouds or the face of Elvis on a piece

of burnt toast. Using magnetoencephalography (MEG), researchers found that the ventral fusiform face area (FFA) in the brain is involved. The FFA has been implicated in detecting faces of people and animals and is also involved in distinguishing animate from inanimate visual stimuli (Kanwisher et al. 1999). This area shows a cortical response 170 ms after we are presented with a human face and shows a similar but slightly earlier activation of 165 ms when seeing objects that resemble faces (Hadjikhani et al. 2009). This suggests that seeing faces is a very early and automatic response and is not something the brain puzzles together after extended cognitive processing. As such, we can assume that responses to robots with a face are early and automatic.

19.3 Design of Social Robots

One aspect that often arises in robot design is that of neoteny, a juvenile appearance that usually evokes a caring response and is generally described as "cute." Young animals, including human children, have a large head, large eyes, chubby cheeks, a small chin, a flat face, a small nose, and relatively short arms and legs. Konrad Lorenz (1982) argued that infantile and juvenile features have a biological function by triggering nurturing responses in adults. We are so keen on neotenous appearances that we breed domesticated animals to retain neotenous features. Many breeds of smaller dogs retain juvenile features, such as a short snout and a relatively large head and large eyes, and consequently are considered cute by most people. The nurturing response is also largely cross-cultural. The same physical features evoke a similar response in people regardless of culture or background. This has been used to good effect by robot designers: if a robot is to be likeable, designers will give it features that evoke a caring response. This not only causes people interacting with the robot to find it cute but also makes them inclined to feel more generous toward any mistakes the robot makes. The opposite seems to hold as well. Robots that have adult, or gerontomorphic, features appear less cute and have less appeal. While there is no research on this yet, it is likely that they are considered more knowledgeable and authoritative, and therefore it makes sense for robot designers to give robots that need to radiate authority or trust an adult appearance (see figure 19.1).

Perhaps the most well-known issue in robot design is that of the *uncanny valley* (figure 19.2). This effect, first hypothesized by Mori in 1970 (Mori et al. 2012), describes the familiarity or appeal of a robot as a function of its human likeness. Mori in his original paper wrote about 親和感 (*shinwa-kan*), which does not translate well into English but is sometimes described as familiarity, appeal, likeability, or affinity. When a robot does not resemble a human, it has low familiarity. This gradually goes up: As human likeness increases, so does familiarity, until the robot is almost humanlike but not quite. At this point familiarity gets knocked back, and when plotted this resembles a sharp dip in the familiarity curve. This is known as the uncanny valley. Androids, robots that have humanlike skin but lack humanlike motions, find themselves firmly in the uncanny valley. You can climb out of the uncanny valley by making a robot that is almost indistinguishable from a person. Note that the uncanny valley effect is more pronounced when the robot is moving: the familiarity or eeriness of the robot is more exaggerated when the robot is animated. Mori never backed up his hypothesis with data, but later empirical research has shown that the

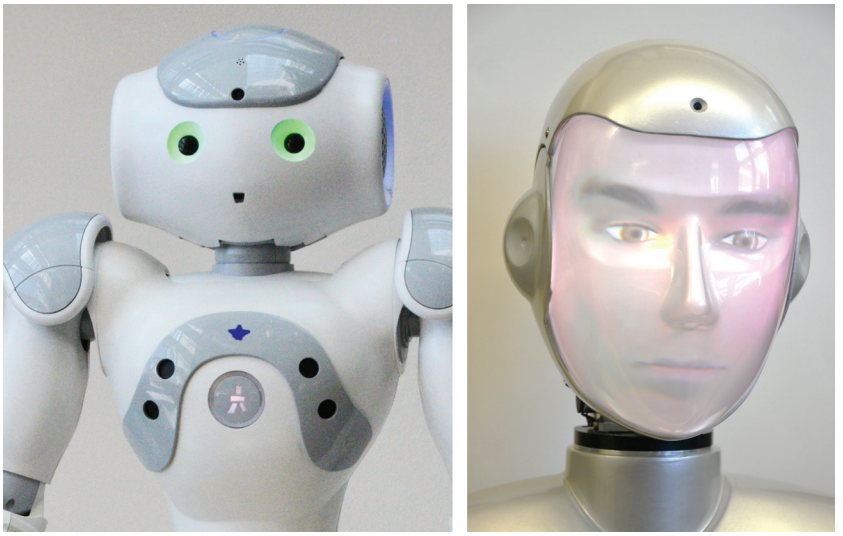

Figure 19.1
A neotenous appearance, characterized by a large forehead, big eyes, a small mouth, and a large head, in robots such as the SoftBank Robotics NAO robot (*left*), make people feel more attracted to them. Robots with adultlike features, such as the Engineered Arts SociBot, which has an adult face (*right*), are likely to be found more authoritative and knowledgeable.

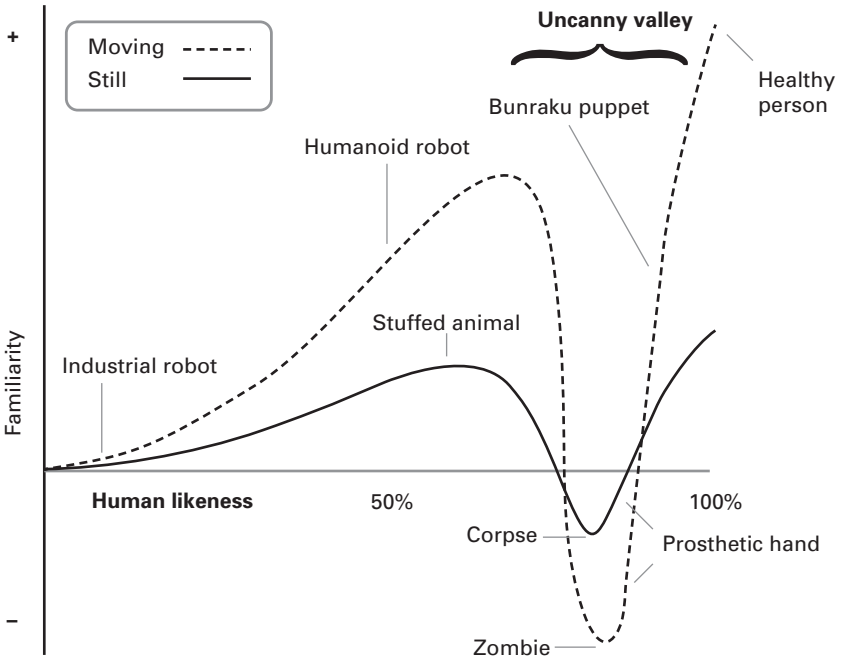

Figure 19.2
A plot showing the uncanny valley, with the famous dip when robots look almost humanlike but repel us because they are not sufficiently humanlike. *Source:* Based on Mori 1970, Wikimedia.

uncanny valley is indeed real (MacDorman and Ishiguro 2006; MacDorman and Chattopadhyay 2016).

Rosenthal-von der Pütten et al. (2019) studied the neural mechanisms underlying human responses to artificial agents and, specifically, the uncanny valley response. They suggest that the uncanny valley requires a neural system that derives human likeness from sensory cues followed by a downstream system that integrates these signals into a nonlinear value function representing the uncanny valley response curve. Using functional magnetic resonance imaging (fMRI), they investigated the neural activity of people when observing people and artificial agents, including robots, while making rated responses or expressing a preference for stimuli. They found that the ventromedial prefrontal cortex encoded a representation of the uncanny valley, in which the subjective likability of artificial agents was a nonlinear function of human likeness. Functionally connected areas in the brain encoded critical inputs for signals: the temporoparietal junction (TPJ) encoded a linear human likeness continuum. The TPJ was also found to be active in detecting agency (Mar et al. 2007), belief attribution, and learning from others (Rosenthal-von der Pütten et al. 2019). In addition, nonlinear representations of human likeness found in the dorsomedial prefrontal cortex (DMPFC) and fusiform gyrus (FFG) emphasized a human-nonhuman distinction. The DMPFC is known to show activity when attributing mental states to others or when assessing performance of others or of the self (Rosenthal-von der Pütten et al. 2019), while the FFG is implicated in distinguishing animate from inanimate stimuli (Chaminade et al. 2010). Activation in the amygdala, which in humans is implicated in the formation and storage of memories associated with emotional events, was found to predict a negative response to artificial agents. As such, the brain seems to have a direct neural representation of the uncanny valley, or rather the uncanny valley can be explained by brain processes that are universal to all people.

If the same neural mechanisms implicated in assessing people, people's behavior, and the agency of stimuli are also active when we perceive robots, then this might help us design more effective robots. Generally, what makes people appealing will make robots appealing, and only cultural conditioning and habituation are likely to change the initial, and often automatic, responses we have to robots.

When discussing the uncanny valley, one cannot escape mentioning androids and perhaps their more famous ilk, the Geminoids. A Geminoid—a contraction of Gemini (meaning "twins" in Latin) and android—is modeled after a human being and as such is their robotic doppelgänger. Hiroshi Ishiguro was the first to build Geminoids, and the various models that have been built—including ones of himself, his daughter, and a Japanese news anchor—have been the subject of academic study into the uncanny valley effect. These studies showed that the uncanny valley effect is sometimes not there or cannot be explained by relying on appearance alone. Bartneck et al. (2009) had people briefly interact with Hiroshi Ishiguro or with his Geminoid. While participants could clearly distinguish an android from a human, and unsurprisingly found the human to be more humanlike, the android was not liked less, which goes against Mori's prediction. This result and others suggest that the uncanny valley is a multidimensional phenomenon and that the two-dimensional plot of figure 19.2 should be revised. Instead the effect is caused by a mismatch between different aspects of the robot: a robot that appears human but moves like a robot causes tension in the observer, which leads to an eerie appearance (Moore 2012).

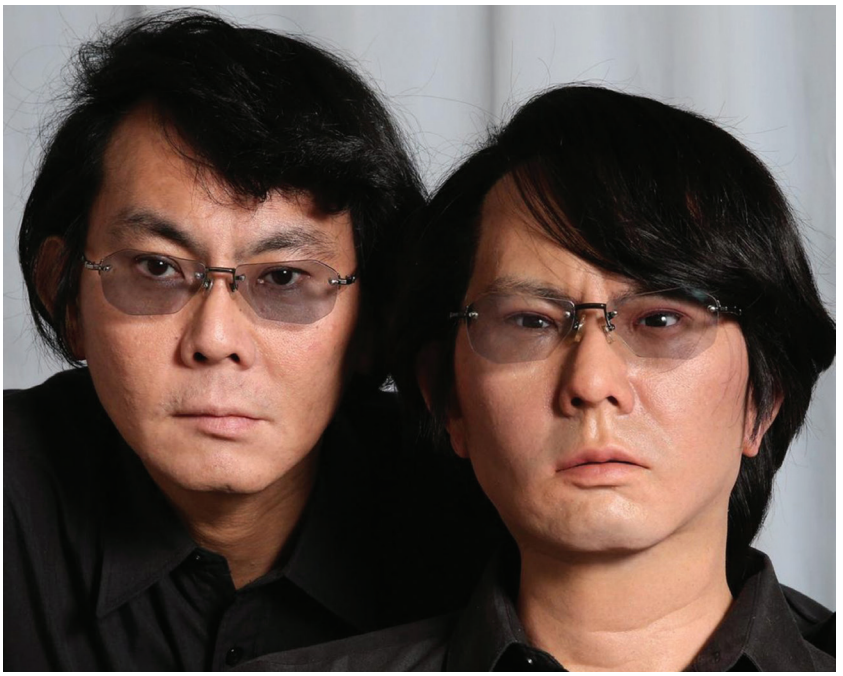

Figure 19.3
Hiroshi Ishiguro and his Geminoid, a robot replica used to study people's responses to lifelike robots. *Source:* Osaka University, Intelligent Robotics Laboratory.

19.4 Verbal Interaction

Social robots will often be addressed using language. Even robots that are not humanlike in appearance, such as animallike robots, are often addressed using speech. Depending on the robot's appearance, people might expect a coherent linguistic response. We don't expect a robot dinosaur to talk back, but we do have expectations of humanoid robots and are invariably somewhat disappointed when those expectations are not met.

In addition, language is most likely to be the most natural and therefore intuitive way to interact with robots. But despite the use of language seeming effortless to us, verbal interaction between people and robots is still a formidable challenge. The typical approach in building natural language interaction (NLI) has been to cut up the problem into several components: speech recognition, dialogue management, language generation, and speech production. And while progress is being made in each of these, unconstrained natural language interaction is still well beyond our technical grasp. Speech recognition, using deep neural networks trained on large sets of annotated speech, now performs better than human transcribers for English spoken by adults (e.g., Xiong et al. 2018). Speech production is almost indistinguishable from human speech for the reading of text with neutral prosody (van den Oord et al. 2016). The developments in speech recognition and speech production have led to a raft of novel applications. Prime examples are the digital assistants, such as Amazon's Alexa or Apple's Siri assistants, that can act on spoken instructions and respond using speech. But these assistants are still very much limited in their func-

tionality, as are most spoken NLI applications. They can take short phrases and take the user through a turn-based dialogue to fill in slots, but they cannot engage in unconstrained dialogue. They do struggle with pragmatic language use—that is, the social language that we use in our daily interactions with others, from the short utterances such as "yup," "sure," or "dunno" that keep linguistic interaction flowing to the extensive reliance on contextual cues to interpret and produce linguistic utterances.

When comparing artificial linguistic interaction systems to language processing in the human brain, it is clear that the two are far apart on several levels. At a fundamental level, language in computers is meaningless to the computer. A chatbot can utter phrases about feelings or the weather, but it does not really understand what it is talking about. It has never experienced feelings or weather, or any other words for that matter. The words that a chatbot uses are not *grounded*. Grounding happens when words and linguistic expressions are experienced and from that become meaningful. The word "chair" only becomes meaningful when a computer or robot has an experiential sensation of a chair by seeing a chair through its camera, by feeling a chair through tactile sensors, or by understanding the function of a chair.

There have been some interesting developments in statistical language processing, where algorithms are used to build models of a language by analyzing large corpora of text. The earliest such algorithms built cooccurrence statistics of words, basically counting which words appeared near others in texts. A distance measure is used to report which words are closer in meaning and which are not. One such technique, latent semantic analysis (LSA), can tell that "king" and "queen" are closely related and that "king" and "lemon" are not (Landauer et al. 1998). New neural network-based approaches take statistical cooccurrence further by learning long-distance dependencies between words. The most recent solutions use recurrent neural networks. At the time of writing, the most notable model is the generative pretrained transformer 3, or GPT-3, but given the arms race between large corporations to outperform each other's language models, the GPT-3 will soon be superseded. The GPT-3 uses transformer networks and was trained on hundreds of billions of words. It was tasked with learning to predict the next word in a sentence and by doing so built a model not only of the English language but also of programming languages (Brown et al. 2020). The GPT-3 seems to have a firm grasp on semantics. It can not only complete sentences; there are impressive examples of it completing short-story lines starting from only an opening paragraph. It can answer questions and passes tests aimed at assessing the vocabulary skills of children. From a cursory inspection, it would seem that the GPT-3 understands language, as it uses language in a very coherent way. However, while the GPT-3 can tell you who the president of the United States is, it would not be able to recognize the president in a photo. The reason, of course, is that the GPT-3, and all other text-based natural language processing systems, are completely text based: the words they use are not *grounded*.

The contrast with human cognition could not be greater: all the words and linguistic constructions we use are grounded in a sensory reality (Harnad 1990). Many have argued that robots should do the same if they are to interact with people in a way in which our exchanges are meaningful (Cangelosi et al. 2002). A robot without grounded linguistic symbols can seem to know the "color of grass," but if it is not able to tie the visual perception of green and grass together, together with all the other memories and cultural agreements on language, human-robot conversation is likely to remain fairly limited.

Another challenge, especially in the context of cognitive robotics, is that language in the human brain is rather poorly understood. We can prod the linguistic brain through behaviorist experiments—for example, by measuring response times to words, which gives us an insight into how words and their meaning might be represented in the brain. Or we sometimes get intriguing views into the linguistic brain through patients who have suffered brain injuries. Important brain regions implicated in language processing and production, such as Broca's and Wernicke's areas, were discovered after studying patients with lesions to those areas. We also discovered that language is to some extent processed in the right hemisphere, after studying patients who had both hemispheres separated by cutting the *corpus callosum*, the part of the brain connecting both hemispheres, but were still able to interpret words shown to only the right visual field.

But even modern brain-imaging techniques have shed relatively little light on how language is processed (Dronkers et al. 2004), represented (Hagoort 2005), and produced in the brain (Levelt 2001) and certainly not to an extent in which insights from cognitive neuroscience would enable us to build better natural language interaction systems. If there is perhaps one valuable lesson, it is that language is not compartmentalized. Instead language seems to permeate the entire brain, with some clear loci for more specific language functions. Artificial NLP, on the other hand, is compartmentalized into components such as speech recognition, language interpretation, dialogue processing, language generation, and speech production while ignoring elements often essential to linguistic communication. Most importantly, the multimodal and nonverbal aspects of communication are largely ignored, and artificial NLP is therefore rather impoverished. Two examples should make this clear: prosody and priming. Prosody is ignored in NLP, although the meaning of a spoken utterance can be completely changed through prosody. Just think of the many ways in which "I'm not at all angry" can be expressed and how the meaning of such a short sentence can swing between joking, furious, irritated, or sad. Human linguistic perception and production is fine-tuned for this, but it remains firmly outside the grasp of artificial speech recognition and production.

Priming is the effect whereby one stimulus influences the response to a later stimulus. For example, asking, "What do cows drink?" often results in people answering "milk" instead of "water" (Rose et al. 2015). Language in the brain is organized as an associative network, with sounds, words (or lemmas), and meaning connected in networks (Collins and Loftus 1975; Levelt 2001). Statistical methods of language modeling, such as hidden Markov models or long short-term memory networks, indispensable in speech recognition and machine translation, explicitly learn statistical associations between phonemes and words. Priming is a very important mechanism both in the brain and in these artificial models: the presentation of a word or phoneme primes, or rather predicts, the next most probable word or phoneme. In the brain, priming is multimodal (Wood et al. 2012), but in NLP the priming only happens within the phonetic or lexical domain, thereby cutting NLP off from modalities that the human brain relies upon to disambiguate and enrich language.

19.5 Nonverbal Interaction

Most content of a natural interaction is contained in its nonverbal aspects. Of course, written text contains very little nonverbal communication (apart from the occasional emoticon) and

seems to work well at conveying information. But spoken language, and specifically language spoken in the presence of others, relies heavily on nonverbal elements. The division of labor between verbal and nonverbal is contested. A widely cited statement is that of Mehrabian (1972), which claims that 55 percent of communication is contained in body language, 38 percent in tone of voice, and only 7 percent in the words spoken. While the exact ratio is up for debate, the fact that verbal communication only accounts for a fraction of communication should point out the flaws in our current efforts in building HRI. For historical reasons most of our technical efforts have been on creating verbal or text-based linguistic interactions while at the same time ignoring nonverbal aspects of interaction. And if we did study nonverbal interaction, we studied it in isolation from other communication channels.

Emotion is a textbook example of this: Due to technical and resource limitations, the first studies of emotion used photographs of facial expressions. Paul Ekman, in his effort to show that some emotions are universal, took a number of photographs of himself and others showing extreme emotions, such as happiness or anger. He indeed confirmed that these emotions are universally recognized and, building on this work, argued that there are at least six or seven basic emotions (Ekman 1972, 1992). Ekman built on a tradition started by Darwin (1872) of using photographs of faces to study emotions, and ever since the discussion of emotions has been dominated by a focus on facial expressions. Nevertheless, faces only show extreme emotions, and emotion is much more likely to be gleaned from context and other body cues (Kappas 2003). In a striking experiment, it was shown that the body posture of tennis players, rather than their facial expressions, showed whether they had won or lost a point, convincingly demonstrating that the face is not necessarily a window to the soul, or to emotions in this case (Aviezer et al. 2012).

Just as with anthropomorphization, the human brain is ever eager to interpret nonverbal signals as meaningful. The clicks, beeps, and whirrs that R2-D2, one of the robot leads from the *Star Wars* series, emits are never interpreted as background noise on the soundtrack of the film but are interpreted as meaningful and relevant by the cinema audience. These clicks and beeps, or nonlinguistic utterances (NLU), can be used to add a nonlinguistic communication channel to robots, complementing language or even short-cutting the need for language. NLUs are interpreted as meaningful by children and adults and can be used to communicate the emotional state of the robot (Read and Belpaeme 2014; see figure 19.4).

Further analysis showed how NLUs are interpreted categorically: if people are asked to interpret an NLU as an emotion, then their interpretation is being drawn to one of only a handful of basic emotions such as happiness, anger, surprise, or fear (Read and Belpaeme 2016). Categorical perception is a fundamental property of perception and is instrumental in interpreting perceptual stimuli. The human brain interprets sensory perception as belonging to a limited number of conceptual states. For example, speech sounds are interpreted as belonging to only a distinct number of phonemes. If hearing a speech continuum in which the amount of voicing is changed gradually, from not at all in "p" in /pa/ to fully voiced in "b" in /ba/, then the perception will be drawn toward known vowels, either "pa" or "ba" but nothing in between. It is surprising that the cognitive mechanisms used to interpret human-human verbal and nonverbal communication are still at work when we are interpreting robotic communicative signals.

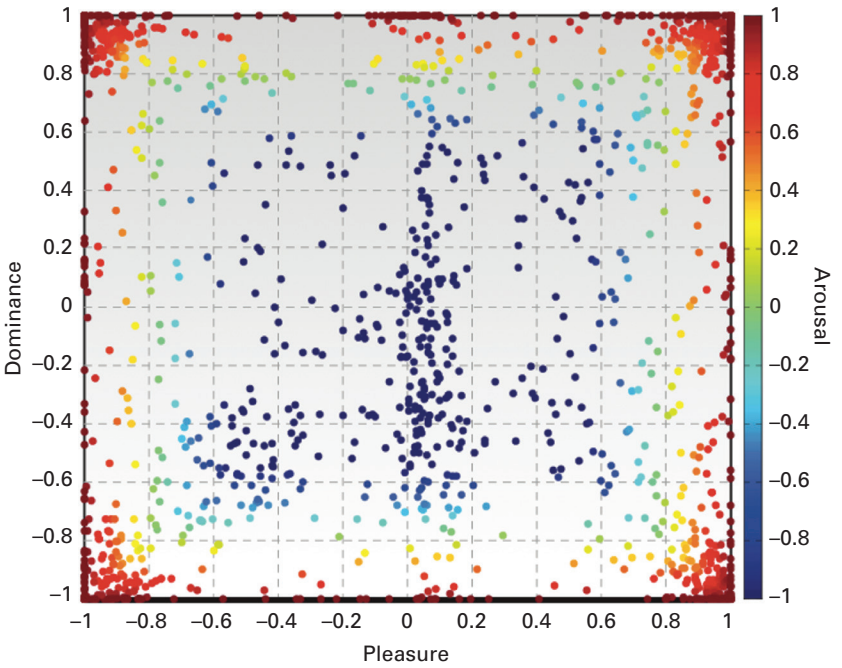

Figure 19.4
Random robot sounds, a concatenation of clicks and beeps, were played to children between six and eight years old. The children were asked to show which emotion the robot was displaying by recreating the emotion on a digital face. These responses were then mapped to a 3D emotion space. Instead of responses being uniformly scattered over the emotion space plot, the children's responses clustered together near basic emotion. This suggests that robot sounds are interpreted as humanlike emotions and that this process is categorical. *Source:* From Read and Belpaeme 2012, 2016.

The combination of verbal and nonverbal interaction, often referred to as multimodal communication in technical parlance, is perhaps the biggest challenge in HRI. One of the reasons for this is that a divide-and-conquer approach, in which a problem is divided up into smaller problems, each to be solved on their own before being recombined to form a total solution, does not seem promising when it comes to building multimodal HRI. In human cognition, multimodal interaction is a complex activity to which all cognitive faculties contribute without clear division, sequence, or hierarchy. For example, hearing a verb (such as "kick") activates the corresponding action in the motor cortex (activity when kicking or thinking about kicking; Pulvermüller 1999), and hearing a naturalistic sound (such as a dog's "woof") and spoken words (/dɔg/) 346 ms before a picture search task led to faster visual detection of the picture of a dog from between distractors (Chen and Spence 2011). It is very likely that the cognitive organization of human interaction will need to be reflected in some way when building HRI. The current separation of processing, with separate components such as speech recognition, dialogue, text to speech, emotion recognition, facial expressions, gesture production, or prosody is artificial and does not have the tight and dynamic coupling that is likely to be necessary for natural HRI.

19.6 Applications

A better understanding of the cognitive mechanisms involved in HRI would surely allow us to build better robots, better interactions, and the best applications. For now, the design of robots and interactions has relied a lot on the gut feeling of designers and engineers and to a lesser extent on theory. However, as soon as HRI is used for applications, an improved understanding of the responses of the human brain to robots might be essential.

Social robots can be used to entertain, persuade, and inform. The strong social character of robots lends itself well to establishing a social bond, and this can be used in diverse applications, such as retail, education, or therapy.

Robots show potential in education. When compared to screen-based learning technologies, such as educational software on computers or tablets, robots tend to have better outcomes. This can be explained by the explicit and tangible social character of the robots, which leads to both improved attitudes toward learning and better learning outcomes. In a metareview (Belpaeme et al. 2018), papers comparing tutoring robots against an alternative, such as educational software or an on-screen avatar, showed that the mean cognitive outcome effect size (Cohen's d) of robot tutoring is 0.70 (95 percent confidence interval (CI), 0.66 to 0.75), which compares favorably to what human tutors can achieve: human tutors achieve an outcome effect size of $d = 0.79$ (Vanlehn 2011). While robot tutors do show promise, designing a robot tutor still is challenging. Robots can be used to tutor restricted domains, such as simple math exercises, but little is known about how to design robot tutors that tackle harder learning challenges. One such challenge is language: the current school-based teaching of a second language relies a great deal on class-based learning of vocabulary and grammar with little to no attention to language use and interaction. This is far removed from how a first language is seemingly effortlessly acquired through interacting with parents, siblings, and peers. The main reason why school-based language learning is so different is that the teacher cannot engage in interaction on an individual basis with all pupils in the classroom. And this is where robots show considerable promise: a robot has the time and infinite patience to interact with those learning a target language. A robot probably also has a better accent than the teacher and can personalize its tutoring to the learner.

Vogt et al. (2019) reported on a large-scale study in which a language-tutoring robot helped young children learn the words and grammar of a second language (see figure 19.5). They used a NAO robot to teach English to five-to-six-year-olds in the Netherlands. Children learned not only nouns ("giraffe" or "boy") but also words used in numeracy (counting words or quantities, such as "more" or "fewer") and spatial language (such as "behind," "in front," and "next to"). The robot tutored the children over seven lessons, introducing six new words during every lesson. The study was used not only to establish whether the robot would be better than only a tablet but also to see whether a robot using gestures to accentuate the words would be a better language tutor. It was divided over four study conditions (a control condition receiving no tutoring, a tablet-only condition, a robot without gestures condition, and a robot with gestures condition), and 208 children took part. While the children did learn English, no significant difference could be found between the learning outcomes: children did not learn more from a robot, whether it was using gestures or not, than from a tablet alone. While there are demonstrations of robots being very effective tutors in narrow domains, the

Figure 19.5
A child learning a second language with the support of a social robot.

benefits of using robots in more complex domains, such as second-language tutoring, are harder won. Robots have been shown to be effective in tutoring vocabulary (van den Berghe et al. 2019), but a more complex use of language probably requires a more complex HRI. A better understanding of how children and adults learn, and how robots can have an impact on this process, will be necessary. It is likely that the social and physical presence of robots is a strong influence on the learning process, but without more open-ended natural interaction, the use of robot tutors is likely to be limited to narrow and closed domains, such as math exercises or vocabulary.

Another application of HRI in which robots are likely to have a significant impact in the future is therapy (Belpaeme et al. 2013). In the last two decades, robotics has been promoted as a promising new technology in autism spectrum disorder (ASD) therapy (Scassellati, Admoni, and Matarić 2012; Thill et al. 2012), and while many supportive case studies exist, there has been a dearth of quantitative empirical evidence about the efficacy of robot therapy (Diehl et al. 2012; Pennisi et al. 2016) that only recently is being resolved. The effect of robots and their behavior on people with ASD is only being studied through the lens of psychological therapy, with little consideration for the cognitive processes involved in the perception of and interaction with robots. It is very likely that a better understanding of the neuropsychology and cognition involved in HRI will allow us to build more effective HRI.

19.7 Conclusion

The relation between human cognition and HRI has largely been explored at the behavioral level. Recently, brain-imaging techniques and response time experiments have given us a

view on how the brain responds to robot stimuli and interactions with robots. All data seem to suggest that interaction with robots relies on the very same social cognitive mechanisms and neural pathways that are also active when we interact with people. This in itself is not very surprising: the brain just generalizes, and our social cognition spills over to nonhuman agents, be they pets or robots. What is more surprising is that our brain readily interprets robotic behaviors, robot forms, and robot noises for which our brain certainly did not evolve. Of course, the nonlinguistic utterances of fictional robots and toy robots have been designed to be interpretable, but even odd combinations—such as a robot vacuum cleaner with a wagging tail (Singh and Young 2012)—remain legible and socially meaningful to us, showing that the human brain really is a most gregarious social interpreter. Understanding how it accomplishes that is likely to lead to a more efficient design of new forms and behavior in HRI.

Additional Reading and Resources

- A classic survey of early approaches to HRI: Goodrich, Michael A., and Alan C. Schultz. 2007. "Human-Robot interaction: A Survey." *Foundations and Trends in Human-Computer Interaction* 1 (3): 203–275.
- A recent, comprehensive volume on HRI: Bartneck, Christoph, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic. 2020. *Human-Robot Interaction: An Introduction*. Cambridge: Cambridge University Press.
- A recent collection on research methods in HRI: Jost, Céline, Brigitte Le Pévédic, Tony Belpaeme, Cindy Bethel, Dimitrios Chrysostomou, Nigel Crook, Marine Grandgeorge, and Nicole Mirnig, eds. 2020. *Human-Robot Interaction: Evaluation Methods and Their Standardization*. Vol. 12. Berlin: Springer.
- A time line of HRI, podcasts on HRI, and additional material accompanying Bartneck et al. (2020): https://www.human-robot-interaction.org/.
- The portal link to the flagship HRI conference in the field and resources on HRI: http://humanrobotinteraction.org/.
- A one-hour video introduction to HRI and social robotics: https://www.youtube.com/watch?v=Lpp1FjkOyN4.

References

Aviezer, Hillel, Yaacov Trope, and Alexander Todorov. 2012. "Body Cues, Not Facial Expressions, Discriminate between Intense Positive and Negative Emotions." *Science* 338 (6111): 1225–1229.

Bartneck, Christoph, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic. 2020. *Human-Robot Interaction: An Introduction*. Cambridge: Cambridge University Press.

Bartneck, Christoph, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. "My Robotic Doppelgänger—a Critical Look at the Uncanny Valley." In *The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 269–276. New York: IEEE.

Belpaeme, Tony, Paul Baxter, Robin Read, Rachel Wood, Heriberto Cuayáhuitl, Bernd Kiefer, Stefania Racioppa, et al. 2013. "Multimodal Child-Robot Interaction: Building Social Bonds." *Journal of Human-Robot Interaction* 1 (2): 33–53.

Belpaeme, Tony, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. "Social Robots for Education: A Review." *Science Robotics* 3 (21).

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." ArXiv preprint: 2005.14165.

Cangelosi, Angelo, A. Greco, and S. Harnad. 2002. "Symbol Grounding and the Symbolic Theft Hypothesis." In *Simulating the Evolution of Language*, edited by A. Cangelosi and D. Parisi, 191–210. London: Springer.

Chaminade, Thierry, Massimiliano Zecca, Sarah-Jayne Blakemore, Atsuo Takanishi, Chris D. Frith, Silvestro Micera, Paolo Dario, Giacomo Rizzolatti, Vittorio Gallese, and Maria Alessandra Umiltà. 2010. "Brain Response to a Humanoid Robot in Areas Implicated in the Perception of Human Emotional Gestures." *PLoS One* 5 (7): e11577.

Chen, Yi-Chuan, and Charles Spence. 2011. "Crossmodal Semantic Priming by Naturalistic Sounds and Spoken Words Enhances Visual Sensitivity." *Journal of Experimental Psychology: Human Perception and Performance* 37 (5): 1554.

Collins, Allan M., and Elizabeth F. Loftus. 1975. "A Spreading-Activation Theory of Semantic Processing." *Psychological Review* 82 (6): 407.

Darwin, C. 1872. *The Expression of the Emotions in Man and Animals*. London: John Murray.

Dennett, Daniel C. 1996. *The Intentional Stance*. 6th ed. Cambridge, MA: MIT Press.

Diehl, J. J., Schmitt, L. M., Villano, M., and Crowell, C. R. 2012. "The Clinical Use of Robots for Individuals with Autism Spectrum Disorders: A Critical Review." *Research in Autism Spectrum Disorders* 6 (1): 249–262.

Dronkers, Nina F., David P. Wilkins, Robert D. Van Valin Jr., Brenda B. Redfern, and Jeri J. Jaeger. 2004. "Lesion Analysis of the Brain Areas Involved in Language Comprehension." *Cognition* 92 (1–2): 145–177.

Ekman, Paul. 1972. "Universals and Cultural Differences in Facial Expressions of Emotions." In *Nebraska Symposium on Motivation*, edited by J. Cole, 207–282. Lincoln: University of Nebraska Press.

Ekman, Paul. 1992. "An Argument for Basic Emotions." *Cognition and Emotion* 6 (3–4): 169–200.

Goodrich, Michael A., Brian Pendleton, P. B. Sujit, and José Pinto. 2011. "Toward Human Interaction with Bio-inspired Robot Teams." In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, 2859–2864. New York: IEEE.

Hadjikhani, Nouchine, Kestutis Kveraga, Paulami Naik, and Seppo P. Ahlfors. 2009. "Early (N170) Activation of Face-Specific Cortex by Face-Like Objects." *Neuroreport* 20 (4): 403.

Hagoort, Peter. 2005. "On Broca, Brain, and Binding: A New Framework." *Trends in Cognitive Sciences* 9 (9): 416–423.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–346.

Heider, Fritz, and Marianne Simmel. 1944. "An Experimental Study of Apparent Behavior." *American Journal of Psychology* 57 (2): 243–259.

Kanwisher, Nancy, Damian Stanley, and Alison Harris. 1999. "The Fusiform Face Area Is Selective for Faces Not Animals." *Neuroreport* 10 (1): 183–187.

Kappas, Arvid. 2003. "What Facial Activity Can and Cannot Tell Us about Emotions." In *The Human Face*, 215–234. Boston: Springer.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25 (2–3): 259–284.

Levelt, Willem J. M. 2001. "Spoken Word Production: A Theory of Lexical Access." *Proceedings of the National Academy of Sciences* 98 (23): 13464–13471.

Lorenz, Konrad. 1982. *The Foundations of Ethology: The Principal Ideas and Discoveries in Animal Behavior*. New York: Simon and Schuster.

MacDorman, Karl F., and Debaleena Chattopadhyay. 2016. "Reducing Consistency in Human Realism Increases the Uncanny Valley Effect; Increasing Category Uncertainty Does Not." *Cognition* 146:190–205.

MacDorman, Karl F., and Hiroshi Ishiguro. 2006. "The Uncanny Advantage of Using Androids in Cognitive and Social Science Research." *Interaction Studies* 7 (3): 297–337.

Mar, Raymond A., William M. Kelley, Todd F. Heatherton, and C. Neil Macrae. 2007. "Detecting Agency from the Biological Motion of Veridical vs Animated Agents." *Social Cognitive and Affective Neuroscience* 2 (3): 199–205.

Marcus, Aaron, Masaaki Kurosu, Xiaojuan Ma, and Ayako Hashizume. 2017. *Cuteness Engineering: Designing Adorable Products and Services*. Berlin: Springer.

Mehrabian, Albert. 1972. *Nonverbal Communication*. New York: Routledge.

Moore, Roger K. 2012. "A Bayesian Explanation of the 'Uncanny Valley' Effect and Related Psychological Phenomena." *Scientific Reports* 2 (2): 864.

Mori, Masahiro, Karl F. MacDorman, and Norri Kageki. 2012. "The Uncanny Valley." *IEEE Robotics and Automation Magazine* 19 (2): 98–100.

Pennisi, Paola, Alessandro Tonacci, Gennaro Tartarisco, Lucia Billeci, Liliana Ruta, Sebastiano Gangemi, and Giovanni Pioggia. 2016. "Autism and Social Robotics: A Systematic Review." *Autism Research* 9 (2): 165–183.

Pulvermüller, Friedemann. 1999. "Words in the Brain's Language." *Behavioral and Brain Sciences* 22 (2): 253–279.

Read, Robin, and Tony Belpaeme. 2012. "How to Use Non-linguistic Utterances to Convey Emotion in Child-Robot Interaction." In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction*, 219–220. New York: IEEE.

Read, Robin, and Tony Belpaeme. 2014. "Situational Context Directs How People Affectively Interpret Robotic Non-linguistic Utterances." In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction*, 41–48. New York: IEEE.

Read, Robin, and Tony Belpaeme. 2016. "People Interpret Robotic Non-linguistic Utterances Categorically." *International Journal of Social Robotics* 8 (1): 31–50.

Reeves, Byron, and Clifford Ivar Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Rose, Sebastian Benjamin, Katharina Spalek, and Rasha Abdel Rahman. 2015. "Listening to Puns Elicits the Co-activation of Alternative Homophone Meanings during Language Production." *PLoS One* 10 (6): e0130853.

Rosenthal-von der Pütten, Astrid, Nicole Krämer, Stefan Maderwald, Matthias Brand, and Fabian Grabenhorst. 2019. "Neural Mechanisms for Accepting and Rejecting Artificial Social Partners in the Uncanny Valley." *Journal of Neuroscience* 39 (33): 6555–6570.

Scassellati, Brian, Henny Admoni, and Maja Matarić. 2012. "Robots for Use in Autism Research." *Annual Review of Biomedical Engineering* 14:275–294.

Singh, Ashish, and James E. Young. 2012. "Animal-Inspired Human-Robot Interaction: A Robotic Tail for Communicating State." In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction*, 237–238. New York: IEEE.

Thill, Serge, Cristina A. Pop, Tony Belpaeme, Tom Ziemke, and Bram Vanderborght. 2012. "Robot-Assisted Therapy for Autism Spectrum Disorders with (Partially) Autonomous Control: Challenges and Outlook." *Paladyn* 3 (4): 209–217.

Thomas, Frank, and Ollie Johnston. 1995. *The Illusion of Life: Disney Animation*. New York: Hyperion.

van den Berghe, Rianne, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne van der Ven, and Paul Leseman. 2019. "Social Robots for Language Learning: A Review." *Review of Educational Research* 89 (2): 259–295.

van den Oord, Aaron, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. "Wavenet: A Generative Model for Raw Audio." ArXiv preprint: 1609.03499.

Vanlehn, Kurt. 2011. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems." *Educational Psychologist* 46 (4): 197–221.

Vogt, Paul, Rianne van den Berghe, Mirjam De Haas, Laura Hoffman, Junko Kanero, Ezgi Mamus, Jean-Marc Montanier, et al. 2019. "Second Language Tutoring Using Social Robots: A Large-Scale Study." In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction*, 497–505. New York: IEEE.

Wood, Rachel, Paul Baxter, and Tony Belpaeme. 2012. "A Review of Long-Term Memory in Natural and Synthetic Systems." *Adaptive Behavior* 20 (2): 81–103.

Xiong, Wayne, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. 2018. "The Microsoft 2017 Conversational Speech Recognition System." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5934–5938. New York: IEEE.