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 towards more natural modes of interaction. These will to a large extent be modelled on how people interact with each other and are comprised of verbal and non-verbal ways of interacting.

HRI is a broad church: at the 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 semi-autonomous drones during a search-and-rescue operation, how can the cognitive load on the operator be optimised to maximise the efficiency of the overall
mission (e.g. Goodrich et al., 2011). On the other end of the spectrum of HRI one finds the 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 as concentrated on it (Bartneck et al., 2020). The holy grail of social HRI of course being the natural and intuitive interaction between people and artificial systems. This on the one hand is a technical effort, with results in social signal processing, artificial intelligence and robotics coming together to create social robots. But social robotics offers an 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”. Not only did we until of late know very little about how people interacted with robots, but our relation and interaction with robots is also continuously evolving. Culture, media, education, context and exposure change our attitudes towards robots and with it the way in which we interact with them. When we meet a robot, several automatic social responses kick in which colour our interaction with the robot, these responses evolved or developed to interact with other human 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 which is known as anthropomorphisation or which
was called the *media equation* by Clifford Nass. We equate media—as computers, printers, mobile phones and of course robots—as if they are human (Byron and Nass, 1996). Everyone has at one time or other muttered at their computer when it crashes or cursed their printer when the paper jams, but the media equation theory takes things a little further by claiming that we not only respond to these media as if they are another person, but that we ascribe personal qualities to media, such as a personality, expertise, and even gender. And we often do so without being aware. The media equation is taken to the extreme in social robots, as the appearance of the robot and its behaviour (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 for the design of social robots, and knowingly or not, designers of social robots and programmers have taken concepts and theories from social psychology into consideration when building robots. Failing to do so usually results in a disappointing human-robot interaction. 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 emotion and intentions has been known for a long time.
Marianne Simmel and Fritz Heider, two psychologists working together in the United States, published an influential paper titled “An Experimental Study of Apparent Behavior” (Heider and Simmel, 1944) in which they describe a simple and elegant experiment: they asked people to describe short video films 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 runtime of the video. 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 self — overinterpreting moving geometric figures. This has been gratefully used by animators and some striking examples exist of very minimalist animation films which show that only very little is needed to nudge our social brain into interpreting simple shapes and movement as having agency (Thomas and Johnston, 1995). If you ever observed a vacuum cleaning robot moving around the room you will probably have been struck by its animal-like appearance, scuttling 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 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 in 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 between animate from inanimate visual stimuli (Kanshiwer et al., 1999). This area shows a cortical response 170ms after we are presented with a human face, and has similar but slightly earlier activation after 165ms 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 that 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 which often comes back in robot design is neoteny, a juvenile appearance which often 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 argued that infantile and juvenile features have a biological function by triggering nurturing responses in adults (Lorenz, 1982). 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, 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 which evoke a caring response. This not only causes people interacting with the robot to find it cute, but also inclines people to feel more generous towards any mistakes made by the robot. The opposite seems to hold as well. Robots that have adult-like 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 who need to radiate authority or trust an adult-like appearance (see figure 19.1).

FIGURE 19.1 HERE

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 (in Mori et al. 2012) describes the familiarity or appeal of a robot in function of its humanlikeness. 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 humanlikeness 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 showed that the Uncanny Valley is indeed real (MacDorman and Ishiguro, 2006; MacDorman and Chattopadhyay, 2016).

FIGURE 19.2 HERE

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 humanlikeness from sensory cues followed by a downstream system that integrates these signals into a nonlinear value function representing
the Uncanny Valley response curve. Using 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, with subjective likability of artificial agents being a nonlinear function of humanlikeness. Functionally connected areas in the brain encoded critical inputs for signals: the temporoparietal junction (TPJ) encoded a linear humanlikeness 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 humanlikeness were 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 from brain processes that are universal between all people.
If the same neural mechanisms implicated in assessing people, people’s behaviour 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 can not escape mentioning androids and perhaps their more famous ilk, the geminoids. A geminoid—a contraction of gemini (meaning twins in Latin) and android—is modelled after a human being and as such is its 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 a Hiroshi Ishiguro or with his geminoid. While participants could clearly distinguish an android from a human, and unsurprisingly found the human to be more human-like, the android was not liked less, which goes against Mori’s prediction. This results and others suggest that the Uncanny Valley is a multidimensional phenomenon and that the 2-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 HERE

19.4 Verbal interaction

Social robots will often be addressed using language. Even robots which are not
humanlike in appearance, such as animal-like 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 in several components: speech recognition,
dialogue management, language generation, and speech production. And while
progress is 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. Xoing et al., 2017). Speech production is almost indistinguishable from human speech for 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. A prime example are the digital assistants, such as Amazon’s Alexa or Apple’s Siri assistants, which can act on spoken instructions and respond using speech. But these assistants are still very much limited in their functionality, 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, i.e. 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 by that they 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 what functions a chair has.

There are some interesting developments in statistical language
processing, where algorithms are used to build models of a language by
analysing large corpora of text. The earliest such algorithms built co-occurrence
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 co-occurrence 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 Pre-trained
Transformer 3 or GPT-3, but given the arms race between large corporations to
outperform eachother’s language models, GPT-3 will soon be superceded.,
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 of the English language but also of programming
languages(Brown et al., 2020). GPT-3 seems to have a firm grasp on semantics.
It not only can complete sentences but there are impressive examples of it
completing short story lines only starting from an opening paragraph. It can
answer questions and passes tests aimed as assessing the vocabulary skills of
cchildren. From a cursory inspection it would seem that GPT-3 understands
language, as it uses language in a very coherent way. However, GPT-3 can tell
you who the president of the United States is, but it would not be able to
recognise the president on a photo. The reason of course is that 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 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 where our exchanges are meaningful (Cangelosi et al., 2002).
A robot without grounded linguistic symbols can seem to know the “colour of
grass”, but if it is not able to tie the visual perception of green and grass together,
together with all 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 behaviourist experiments, through for example
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 suffered brain
injuries. Important brain regions implicated in language processing and
production, such as Broca’s and Wernicke’s area, were discovered after studying patients with lesions to those areas. Or we discovered that language is to some extent also 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 where 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 compartmentalised, instead language seems to permeate the entire brain, with some clear loci for more specific language functions. Artificial NLP on the other hand is compartmentalised in 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 non-verbal 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, while 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 finetuned to 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 organised as an associative network, with sounds, words (or lemmas) and meaning connected in networks (Collins and Loftus, 1975; Levelt, 2001). Statistical methods of language modelling, 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 phonemes 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 which the human brain relies upon to disambiguate and enrich language.

19.5 Non-verbal interaction

Most content of a natural interaction is contained in its non-verbal aspects. Of course, written text contains very little non-verbal 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 non-verbal elements. The division of labour between verbal and non-verbal is contested. A widely cited statement is that of Mehrabian, which claims that 55% of communication is contained in body language, 38% in tone of voice, and only 7% in the words spoken (Mehrabian, 1972). While the exact ratio is up for debate, the fact that verbal communication only accounts for a fraction of communication should point out the flaw in our current efforts in building Human-Robot Interaction. For historical reasons most of our technical efforts have been on creating verbal or text-based linguistic interactions, while at the same time ignoring non-verbal aspects of interaction. And if we did study non-verbal 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 recognised 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) to use photographs of faces to study emotions, 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 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 if they won or lost a point, convincingly showing that the face is not necessarily a window to the soul, or to emotions in this case (Aviezer et al., 2012).

Just as with anthropomorphisation, the human brain is ever eager to interpret non-verbal signals as meaningful. The click, beeps and whirrs that R2D2, one of the robot leads from the Star Wars series, emits are never interpreted as background noise on the sound track of the film, but are interpreted as meaningful and relevant by the cinema audience. These clicks and beeps, or Non-Linguistic Utterances (NLU), can be used to add a non-linguistic communication channel to robots, complementing language or even shortcutting the need for language. NLUs are interpreted as meaningful by children and adults and can be used to communicate emotional state of the robot (Read and Belpaeme, 2014), see Figure 19.4.

FIGURE 19.4 HERE

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 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 towards known vowels, either “pa” or “ba” but nothing in between. It is surprising that cognitive mechanisms used to interpret human-human verbal and non-verbal communication still are at work when we are interpreting robotic communicative signals.

The combination of verbal and non-verbal 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 then 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, the hearing of a verb (such as “kick”) activates the corresponding action in the motor cortex (activity when kicking or thinking about kicking) (Pulvermüller, 1999) or the hearing of a naturalistic sounds (such as a dog’s “woof”) and spoken words (/dog/) 346 ms before a picture search task led to faster visual detection of the picture of a dog from between distractors (Chen and Spence,
It is very likely that the cognitive organisation 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 which 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 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 robot 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 diverse application, 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 towards learning and better learning outcomes. In a meta-review (Belpaeme et al., 2018) papers comparing tutoring robots against an alternative, such as an 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% confidence interval (CI), $0.66$ to $0.75$], which compares favourably to what human tutors can achieve: human tutors achieve an outcome effect size of $d = 0.79$ (VanLehn, 2011). While robot tutors do promise, designing a robot tutor still is challenging. Robots can be used to tutor restricted domains, such as simple math exercises, but little know about how to design robot tutoring that tackle harder learning challenges. One such challenge is language: the current school-based teaching of a second language relies a lot on class-based learning of vocabulary and grammar with little to no attention for 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 a learning in target language. Not only that, it probably also has a better accent than the teacher has and can personalise its tutoring to the learner.

Vogt et al. (2019) report on a large-scale study in which a language tutoring robot helped young children learn words and grammar of a second language
(see Figure 19.5). They used a NAO robot to teach English to 5 to 6 year old in the Netherlands. Children learned nouns (“giraffe” or “boy”), but also words used in numeracy (such as counting words or quantity, such as “more” or “fewer”) and spatial language (such as “behind”, “in front”, “next to”). The robot tutored the children over 7 lessons, introducing six new words every lesson. The study was not only used to establish if the robot would be better than only a tablet, but was also used to see if a robot using gestures to accentuate the words would be a better language tutor. 208 children took part in the study, 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). While the children did learn English, there was no significant difference 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 demonstration of robots being very effective tutors in narrow domains, the benefits from using robots in more complex domains, such as second language tutoring, is harder won. There are some demonstration of robots being effective in tutoring vocabulary (van den Berghe et al., 2019), but more complex use of language probably requires a more complex Human-Robot Interaction. A better understanding of how children and adults learn, and how robots can 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 maths 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 there are many supportive case studies, there has been a dearth of quantitative empirical evidence about the efficacy of robot therapy (Diehl et al., 2012; Pennisi et al., 2016) which only recently is being resolved. The effect of robots and their behaviour on people with Autism Spectrum Disorder (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 human-robot interaction has largely been explored at the behavioural 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 seems to suggest that the 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 generalises and our social cognition spills over to non-human agents, be they pets or robots. What is more surprising is that our brain readily interprets robotic behaviours, robot forms and robot noises for which our brain certainly did not evolve. Of course, the non-linguistic 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. And understanding how it does that is likely to lead to a more efficient design of new forms and behaviour in HRI.

References


Read, Robin, and Tony Belpaeme. "How to use non-linguistic utterances to convey emotion in child-robot interaction." In *2012 7th ACM/IEEE*


Vogt, Paul, Rianne van den Berghe, Mirjam de Haas, Laura Hoffman, Junko Kanero, Ezgi Mamus, Jean-Marc Montanier et al. "Second language tutoring using social robots: a large-scale study." In 2019 14th ACM/IEEE


**ADDITIONAL READING**


FIGURE CAPTIONS

Figure 19.1: A neotenous appearance of robots, such as a large forehead, large eyes, a small mouth, and a large head, make people feel more attracted to robots, such as the Softbank Robotics Nao robot (left). Robots with adult-like features are likely to be found more authoritative and knowledgeable, such the EngineeredArts Socibot showing an adult face (right).

Figure 19.2: the Uncanny Valley, a plot showing familiarity of a robot in function of humanlikeness with the famous dip when robots look almost humanlike, but repel us because they are not sufficiently humanlike (based on Mori, 1970, Karl MacDorman).

Figure 19.3: Hiroshi Ishiguro and his geminoid, a robot replica used to study people’s responses to lifelike robots (Osaka University, Intelligent Robotics Laboratory).

Figure 19.4: Random robot sounds, sounding like a concatenation of clicks and beeps, where played to children between 6 and 8 years old. They 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 human-like emotions and that this process is categorical (Read and Belpaeme, 2012, 2016).

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

FIGURES

19.1
19.4