CellPress
OPEN ACCESS

Trends in
Cognitive Sciences

Series: Machine Behavior

Opinion

Mind Meets Machine: Towards a Cognitive Science of Human–Machine Interactions

Emily S. Cross [1,2,4,*] and Richard Ramsey [3,4]

As robots advance from the pages and screens of science fiction into our homes, hospitals, and schools, they are poised to take on increasingly social roles. Consequently, the need to understand the mechanisms supporting human–machine interactions is becoming increasingly pressing. We introduce a framework for studying the cognitive and brain mechanisms that support human–machine interactions, leveraging advances made in cognitive neuroscience to link different levels of description with relevant theory and methods. We highlight unique features that make this endeavour particularly challenging (and rewarding) for brain and behavioural scientists. Overall, the framework offers a way to study the cognitive science of human–machine interactions that respects the diversity of social machines, individuals' expectations and experiences, and the structure and function of multiple cognitive and brain systems.

Highlights

Although machines designed to socially interact with humans are proliferating, our understanding of the mental processes supporting such interactions remains limited.

The cognitive and brain sciences can make considerable theoretical and methodological contributions to this endeavour.

To date, the modal approach to cognitive science-informed human–robot interaction research has been grounded in social cognition.

We widen the lens through which much of this research has been framed to incorporate diverse and interacting forms of cognition, which emphasise the complexity of understanding human–robot interaction on a mechanistic level.

Our framework seeks to account for the unusual demands on (neuro)cognition presented by engaging with social machines, focusing on machine variety, diverse neurocognitive systems, and unique contextual factors such as media-skewed expectations.

From Social Cognition to Social Machines

Over a decade ago, Microsoft founder Bill Gates prophesied a robotics revolution that would see staggering leaps in the progress and sophistication of robots, and predicted 'a robot in every home' in the near future [1]. Although the ubiquity of household robots has yet to be realised, we are opening our doors to an increasing number of **artificially intelligent machines** (see Glossary). Concurrently, a growing number of robotics start-ups are focusing on developing companion robots for the home or assistance robots to serve in complex, human-interactive contexts, including schools, hospitals, and care homes [2]. As progress towards developing machines that take on increasingly sophisticated social roles continues apace, the cognitive and brain mechanisms that underpin social engagement with these machines remain largely unknown. A greater understanding of the psychological and neurobiological foundations of human interactions with **artificially intelligent social machines** (henceforth referred to as 'social machines') has important implications for the design and programming of socially engaging and collaborative artificial agents. It is equally critical to use this understanding of human–machine interaction to further our knowledge of the flexibility and limits of neurocognitive processes supporting human social behaviour towards both human and artificial agents.

Investigating the cognitive and brain mechanisms that underpin human–machine interactions is a fledgling field that faces several unique challenges. For example, the vast range of artificially intelligent machines, from small handheld devices such as smartphones and thermostats to autonomous petlike robots, such as Paro and MiRo, all the way to life-sized humanoid robots, such as iCub or Pepper, suggests that this endeavour represents a highly variable space (Figure 1). Interactions with different robots are likely, therefore, to involve a range of different mental processes, making a 'one size fits all machines' type of cognition unlikely. Furthermore, even the most sophisticated social robots share features with inanimate objects and rudimentary machines in addition to animate beings, such as animals and humans. Consequently, the typical dividing lines of cognition that carve up the environment on the basis of clear categories are

[1]Department of Cognitive Science, Macquarie University, Sydney, Australia
[2]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, Scotland, UK
[3]Department of Psychology, Macquarie University, Sydney, Australia
[4]Both authors contributed equally to this article

*Correspondence:
emily.cross@mq.edu.au (E.S. Cross).

 https://doi.org/10.1016/j.tics.2020.11.009

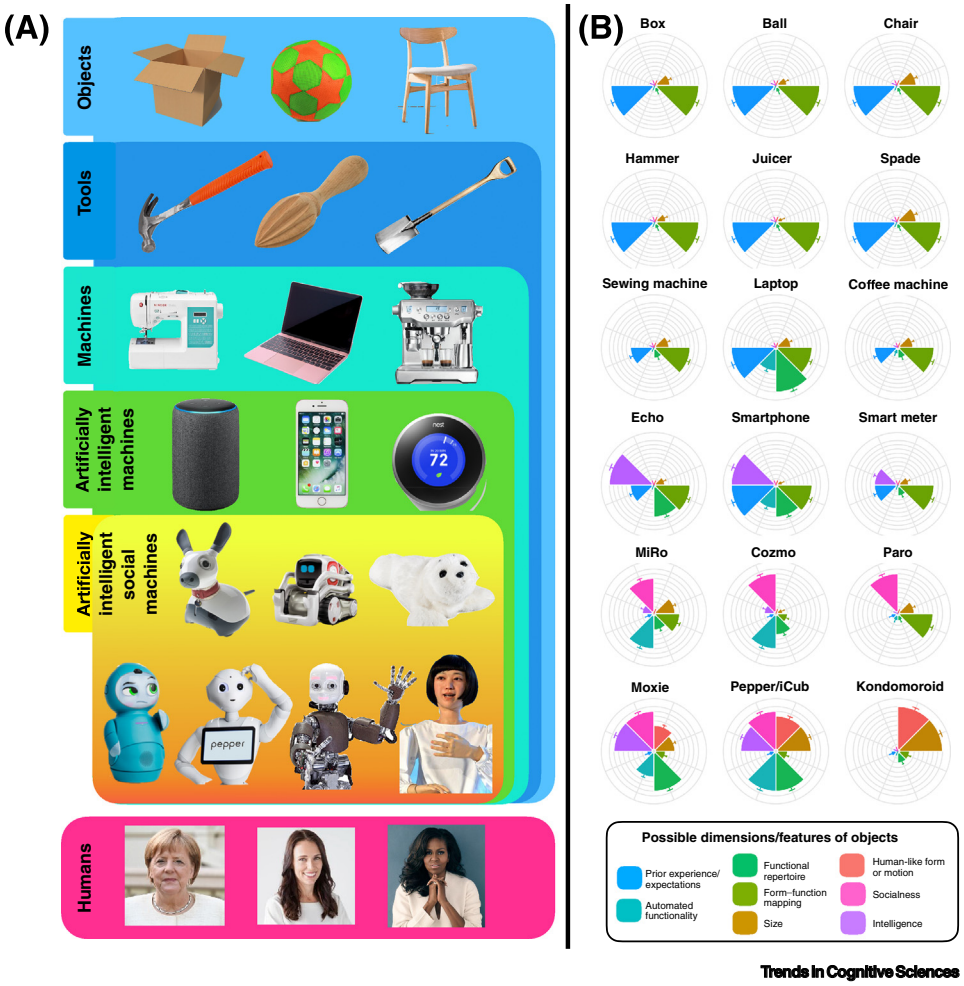

Figure 1. Possible Ways to Codify and Distinguish Objects, Ranging from Chairs to Social Robots. (A) Categorisation of objects based on broadly similar features attempts to group objects into distinct kinds, such as tools, machines, artificially intelligent machines, and artificially intelligent social machines. Naturally, however, the variety of exemplars within each of these rough categories is enormous. (B) An alternative or complementary way to describe objects (including artificial agents) is via a dimensional or feature-mapping approach. Dimensional approaches ignore strict categorisation and instead focus on evaluating and comparing objects across a range of dimensions. Therefore, objects and machines can be more or less similar to each other across multiple dimensions or features. The eight dimensions shown here are given as possible examples to illustrate the value of taking a dimensional approach. We do not intend these dimensions to represent a comprehensive list of relevant features; instead, we use them to spark further debate and research concerning which features may be most relevant in particular contexts and for particular research aims. N.B.: The actual names of the robots shown in A (referred to in the main text) are stated over the corresponding feature map cartoons in B.

broken. This presents a particular challenge when one attempts to understand how the brain constructs mental representations that guide interactions with artificially intelligent machines.

Although the work of engineers, roboticists, behavioural scientists, and philosophers has done much to inform our understanding of **human–robot interaction (HRI)**, here we leverage discoveries made in human cognitive neuroscience to outline a framework for studying human–machine interactions that holds potential for advancing our understanding in new ways. We focus on human interactions with robots that have been designed to take on social roles. As a starting point in this

Glossary

Artificially intelligent machines: sometimes referred to as 'intelligent agents,' these are physically embodied objects (such as a phone or thermostat) that are equipped with technology that enables them to perceive aspects of their environment to help them achieve their programmed goals (such as unlocking a phone by comparing a user's face with stored templates of the user's face, or a smart thermostat illuminating temperature information on its display when it detects a person entering a room).

Artificially socially intelligent machines: building off the core capacities of artificially intelligent machines, machines with artificial social intelligence are designed to either detect and respond to social signals in the environment or detect and respond to signals in the environment in a way that is perceived as social by human users, or some combination of these two possibilities. These are also referred to in the literature as 'social robots' or 'socially assistive robots.' They are referred to as 'social machines' in the main text for simplicity.

Domain-general: processes that operate across different stimulus or task features.

Domain-specific: processes that are tuned to particular stimulus or task features.

Embodied robots: robots that are physically present and co-located in space with human users. Compare these with virtual agents or voice agents, which are not considered embodied according to this definition.

Human–robot interaction (HRI): a multidisciplinary research field seeking to understand, design, and evaluate human interactions with robots. HRI research covers all varieties of HRI, including factory assembly robots, robotic protheses, and drones, whereas the type of HRIs that are the focus of this piece are human interactions with robots designed for social purposes.

Machines: apparatuses using mechanical power and having several parts, each with a defined function and together performing a particular task (definition from the *Oxford English Dictionary*).

Social cognition: cognitive processes that are engaged by thinking about and interacting with other people.

endeavour, we propose an updated framework that broadens the focus of HRI research beyond **social cognition** and that considers (i) the important and distinct roles played by machine variation, (ii) diverse and interacting cognitive and brain systems, and (iii) the unique challenges and opportunities that research into human interactions with social machines presents (Figure 2C).

The primary value of placing HRI research into this broad framework is to better characterise the interplay between **domain-general** information-processing systems, which operate across a

Social robotics: a term encompassing a wide variety of research relating to robots designed to engage humans on a social level, often in companionship or assistance contexts. HRI is one facet of this diverse field.

Turing Test: developed by British computer scientist and mathematician Alan Turing, the Turing Test is a method of inquiry in artificial intelligence where a human user probes whether a computer (or agent) they are speaking with is being controlled by another human user or by artificial intelligence.

Wizard of Oz: describes a methodological approach most often used in laboratory studies of HRI where a robot's behaviours or responses are controlled either partially or fully by a human experimenter, unbeknownst to the human participant.

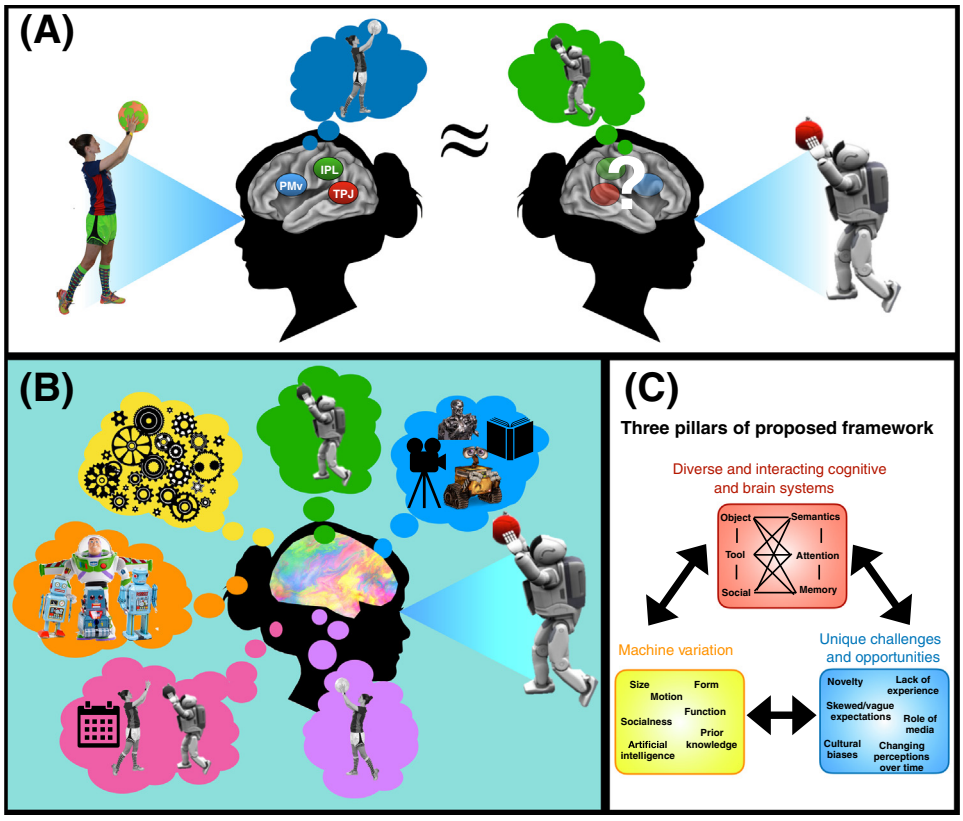

Figure 2. Schematic Depictions of Possible Approaches to Studying Human Interactions with Social Machines. (A) Traditional neurocognitive approaches have focused on using human social cognition as a template upon which to base or compare our understanding of how robots are perceived, and how we interact with them, in social settings. A particular focus has been on engagement of brain regions implicated perceiving, interacting with, and thinking about other people in social contexts, including the ventral premotor cortex (PMv), the inferior parietal lobule (IPL), and the temporoparietal junction (TPJ). (B) An updated approach, called for in the present article, is to move beyond a reliance on (human) social cognition and simply using the self (or another person) as a template for engaging with social machines. Instead, a fuller appreciation of more general aspects of cognition, such as object and tool perception, as well as an individual's expectations (shaped by media, culture, and other factors), prior experience with robots (e.g., toys), amount of time previously spent engaging with the specific robot, and a broadening of focus beyond brain regions implicated in social cognition, should accelerate progress in this endeavour. (C) Schematic summary of the three main pillars of the proposed framework. The framework highlights how to advance understanding of the mechanisms supporting human interactions with social machines via consideration of (i) a diverse set of cognitive and brain systems that span interactions between general and more specialised information-processing streams, (ii) widespread variation in machine features, and (iii) the unusual demands placed on cognitive systems by human–machine interactions that present researchers with unique challenges and opportunities. This framework may be best considered as a platform to encourage researchers to take a broader perspective that includes more diverse viewpoints when studying cognitive and brain systems involved when humans interact with social machines. In the future, this framework should also provide the foundation for developing and testing neurocognitive models that make more specific predictions regarding the human side of human–robot interaction.

range of stimuli, tasks, and contexts, and those that are more **domain-specific** and tied to interactions with other agents, such as humans and social machines. In line with this, broadening the focus of HRI research to include theoretical and empirical perspectives beyond social cognition will be important because, even though considerable progress has been made in understanding social cognition, it would be a mistake to assume that our understanding of what is 'social' about social cognition is clear-cut or finalised [3,4]. Therefore, adopting a broader framework to examine how we perceive and interact with social machines promises to be fruitful because (i) it is more agnostic about the cognitive processes supporting HRIs, and (ii) it could also help us to develop a more nuanced understanding of what, precisely, is 'social' about social cognition.

Considering Machine Variation Jointly Along Dimensional and Categorical Axes

To understand how people might perceive and interact with social machines, such as those the field of **social robotics** aims to develop, it is useful to place such machines within a wider context of entities that we are likely to encounter in daily life. We can think of **machines** as a subclass of objects and tools that are physically present in the world and perform specific tasks. A wealth of cognitive neuroscience research has charted the human brain's response to perceiving and categorising salient stimuli in the environment, including faces, body parts, objects, and tools (e.g., [5–11]; for reviews, see [12,13]). This work generally converges on the finding that object and tool perception engages a core neural network comprising premotor, parietal, lateral occipital, and ventral temporal areas. Cognitive neuroscience research has also demonstrated that social interactions are supported by a distributed neural network spanning perceptual, affective, and regulatory functions, which is partly dissociable from the neural networks underpinning our interactions with objects and tools [14,15]. Therefore, much progress has been made in understanding how we perceive and interact with objects, tools, and animate beings, based on the assumption that such entities are processed by partly dissociable neurocognitive systems [14,16,17].

By contrast, our understanding of how we perceive and categorise artificially intelligent machines, let alone social machines, remains limited for several reasons. First, the cognitive neuroscience of HRI is a new field of inquiry, with a research programme in its infancy [18,19]. Second, compared with divisions between humans, houses, and animals, for example, clear-cut categories are not as easily agreed upon when it comes to machines, especially artificially intelligent social machines. Definitions of what makes a robot 'social' or 'intelligent' are slippery, may vary across people, and are generally difficult to agree upon. With this being said, coarse categorisation is possible to some extent: Machines have automated functions compared with objects and tools, and an objective boundary exists between even the most sophisticated machines and humans (Figure 1A). Nonetheless, the variety of machines within and between loosely defined categories is vast in terms of size, shape, motion profile, and intended use or purpose (Figure 1B). This variation matters because, if neglected, it poses a threat to the tractability of human–machine interaction research (Box 1).

A dimensional or 'feature-mapping' approach may offer a solution to the challenging, and likely fruitless, endeavour of strictly categorising the socialness and intelligence of robots (Figure 1B). A number of research teams have previously touched upon the topic of feature mapping from both HRI [20–23] and social cognition [24,25] perspectives. However, previous proposals have almost exclusively focused on social and psychological aspects of perceiving robots, thus making comparisons between objects, social machines, and humans difficult. By contrast, the type of dimensional approach we propose has a broader scope and considers all objects (including social machines) across a range of relevant dimensions (e.g., form, motion, size, social capabilities).

Our approach is not intended to be prescriptive about which dimensions are most important in a given context, as the development and validation of dimensions demand their own substantial line

Box 1. Respecting the Variety of Machines and Research Goals

Variety in HRI research is not limited to the type of machine; a wide variety of aims also exist. Such aims range from understanding different levels of description such as psychological, cognitive, and neurobiological, as well as work that aims to probe how basic systems operate when people encounter and interact with robots, right up to applications in robotics, such as in the service industry, education, healthcare, and therapy. These aims also include whether a particular robot meets the specific requirement for which it was originally built. Much like recent proposals in computational neuroscience, where a wide variety of aims and objectives also proliferate [89], our argument is that HRI research will likewise benefit from authors being clear about the stated aims of their work. Considerable debate, disagreement, and possibly confusion exist regarding the purpose of computational modelling in neuroscience precisely because computational modellers are pursuing not one goal, but many. We see many parallels between the computational neuroscience domain and HRI research in this regard. Consequently, if HRI researchers were to routinely and explicitly justify the goal or goals of each individual study, as discussed in a recent review paper [72], this should facilitate comparisons across studies, methods, and robotics platforms, thus leading to more rapid progress in HRI as a field.

As one example, important differences exist between research that aims to understand the psychological aspects underpinning HRIs in terms of how individuals think and act toward robots, and research that aims to investigate the underlying cognitive and brain systems supporting HRI. Indeed, understanding how one feels or behaves towards a specific robot (compared with humans or other robots, for example) may not always inform the structure of cognitive and brain function. Likewise, the study of cognitive and brain function may not always yield useful insights into how a person may feel or behave towards a robot. As such, the field of HRI stands to benefit if researchers are clearer regarding the extent to which their aims are to understand how people perceive and interact with robots in terms of psychological variables and/or to make inferences about the structure and function of underlying (neuro)cognitive systems (cf. [70]). Such research aims can, of course, be related but are nonetheless distinct in important ways.

of research. Indeed, the individual features being evaluated might be nonlinear and multidimensional themselves, with their importance varying depending on one's research questions, as well as complex contextual factors, such as how a machine is introduced to a person, the person's age, cultural background, or what they believe about that particular machine (cf. [26–29]). Finally, the dimensions outlined in Figure 1B are simple examples to illustrate the value of a dimensional (compared with a categorical) approach rather than a suggestion that these dimensions should receive privileged status. As such, we anticipate that different research questions may be best served by consideration of different dimensions.

More generally, dimensional approaches have been used successfully in psychopathology, where one-to-one mappings between symptoms and diagnoses are rare, thus limiting the utility of neat categorisations [30–33]. Likewise, person perception is often construed as a multidimensional 'space' that maps across a range of features, including faces, bodies, and traits [34–37]. In social robotics research, combining higher-order broad categorisation with lower-order feature mapping (including nonsocial features) may help establish more accurate expectations regarding machines. For instance, an agent may look human-like but have limited social abilities (e.g., android robots, such as Kondomoroid; Figure 1A) or may have human-like limbs but be unable to replicate biological motion (e.g., the robots Pepper and iCub; Figure 1A). Such feature combinations matter for psychological and brain-based research because an agent may be relatively human-like along one dimension but not at all in other dimensions, which is likely to impact mental processes in (as yet) unknown ways.

Placing a greater emphasis on the variety of machines and their associated features is the first central pillar of the current proposal and has both theoretical and practical implications. In terms of theory development, under a dimensional account, it becomes unsurprising that an artificial agent could engage systems similar to those of a human. This is because such agents were explicitly designed to look or behave in a human-like manner on some level [28,38–41] (Box 2). From an applied perspective, rapid progress continues toward developing social machines for a variety of applications, including domestic service, education, and healthcare [42,43]. Each

Box 2. Cautionary Tales for Cognitive Neuroscience Investigations into Human–Machine Interaction

Like any new field of research, research examining human interactions with social machines can learn from the successes and failures of other research domains [74]. We highlight several cautionary tales that cover feature similarity across agents and theory development. First, in terms of feature similarity across agents, a danger exists that one could take an object (such as a cardboard box), add human features (such as a pair of eyes), and then show similarity on some level between the mechanisms that process our interactions with this box and with other people. It seems difficult for the outcome to be otherwise because, by construction, the object has been given those qualities, and we have known for over 70 years that simple shapes can be perceived to have human-like qualities (e.g., [39]).

The situation is reminiscent of research that has used reaction time paradigms and social stimuli. If you compare a social stimulus (e.g., eye gaze, body orientation) and a nonsocial stimulus (e.g., an arrow), which both have a common feature (a directional cue), then some similarity in processing should be expected in terms of directional cueing. Indeed, using a Posner-style cueing task [90], eye-gaze stimuli that look leftward or rightward cue attention and in a similar manner to arrows [91,92]. Likewise, a body or an arrow that faces left or right also biases the speed of judgments to locations in space [93,94]. Of course, this does not mean that these stimuli are identical. It simply means that they partly share a directional cue. When people and arrows are set in real-world contexts, it becomes easier to see how directional cues from people and arrows may operate differently, such that social cues dominate arrow cues [95,96]. As such, we feel that feature similarity in social cognition research is an important strand of research to consider when designing and interpreting HRI studies. Moreover, this also underscores why research performed with embodied, physically co-located robots is so vital for advancing HRI research compared with work that focuses on static robot images presented on screens [54,80,97], and this echoes calls for increasing use of 'real-life' neuroscience approaches in general [44,98,99].

The second cautionary tale concerns theory development. Consider a prototypical human neuroimaging study that shows relatively more or less engagement (or effective connectivity or pattern similarity) in brain regions during conditions that involve a robot compared with a human. If these findings are not supported by sufficient theory to say why or how they should or should not differ, how should one interpret the results? The often-implicit alternative hypothesis represents a null hypothesis; that is, no overlapping systems should emerge. However, a completely nonoverlapping hypothesis represents a weak theoretical position, because it seems unlikely that such an object would be processed without any similarity to human interactions based on the way such robots are designed to share human features. Without clearer theoretical foundations, therefore, which should facilitate more specific alternative predictions [100], the work is almost impossible to falsify, thus reducing its value.

context is already inspiring the development of a host of different robots brought to market, each designed to perform a particular task or service. This highlights the value and timeliness of clarifying the extent to which neurocognitive mechanisms of HRI generalise across robots and contexts.

Broadening the Theoretical Field of View to Incorporate Diverse and Interacting Neurocognitive Systems

As emphasised in a recent paper [44], most research examining how people perceive social robots has focused almost exclusively on comparing brain and behavioural responses when people encounter robots versus humans (e.g., [28,38,40,45–51]). Built into this approach is a largely implicit assumption that it is useful to consider a robot as a human wearing a tin suit. To some extent, this approach is valid and can add value. If robots have human qualities by design, it seems sensible to use human social interaction as a model system [19,52–57]. For example, using social cognition as a model system might tell us whether perceiving a particular robot's face compared with a human face is associated with activation of overlapping or distinct brain regions [47].

However, although progress has been made using social cognition as the default model, restricting theoretical and empirical focus to this model will fundamentally limit further progress for at least three reasons. First, as discussed earlier (also see Box 2), adding human features to a non-human object and then measuring similarity in a behavioural or neural measure to human interactions can only get us so far. By construction, the object was given human qualities, and therefore it seems almost guaranteed to be processed similarly to a human on some level or to some degree. It remains unclear, however, whether this adds to our understanding of cognitive and brain systems in meaningful ways. Second, robots share qualities with tools and objects,

as well as nonsocial artificial agents and non-human animals [58]. In principle, therefore, research examining object, tool, and animal interactions seems as valuable as social cognition research (cf. [44]). Third, general cognitive and brain functions, such as those associated with aspects of memory, attention, and semantics, have been undervalued in HRI research to date, even though strong arguments exist for considering these in combination with more specialised, category-selective processes [59–61].

We suggest that the shared feature space between social machines and objects holds as much potential to develop an explanatory framework as do the shared features between social machines and humans. Moreover, several advantages come with embracing additional and broader domains of cognition. First, our understanding of object perception, as well as generalized cognitive processes, is more developed than social cognition. For example, in object perception, more comprehensive understanding exists about interactions between visual cortex and anterior brain regions integrating visual object features with top-down factors, such as emotional valence and context (e.g., [62–64]). As a consequence, a greater focus on the similarities, rather than differences, in underlying cognitive architectures supporting object and social perception holds value. It is worth noting that the same observation was made in seminal early social cognition work by Fritz Heider [65], which has been overlooked in recent years in favour of a focus on differences between objects and people. Likewise, memory and attention have been studied for decades, with work spanning multiple species and using a variety of methods. This provides an extensive evidence base upon which to ground new hypotheses and research questions that are also relevant to HRI research [66–69].

Second, increased consideration of how broader and more established research programmes mesh and interact with social cognition can help sharpen the specificity of the inferences being made. For example, given that social cognition is a product of domain-general and domain-specific (social) systems, a key inference concerns stimulus specificity [14]: Are observed effects tied to the socialness of the stimulus in question, or do they reflect a more general mechanism that applies to many categories and features of objects? The same logic applies to studying encounters with social machines. Study designs that compare responses when we perceive or interact with humans compared with robots can lack sufficient stimulus specificity to demonstrate the involvement of specifically social cognitive and brain processes. A consequence is that outcomes can be difficult to interpret in the ways that HRI researchers typically want to interpret them, which is in relation to specifically social brain systems being ‘reused’ during interactions with non-human agents (e.g., [70]). What follows can be a mismatch between the level of specificity inferred and the level of specificity demonstrated by the evidence. An alternative and deflationary position is that general mechanisms of attention (as one example), which are not tied to social stimuli, such as inhibition, alerting, filtering, and orienting, are involved in interactions with humans and machines. Therefore, an important challenge for future theoretical and empirical work into HRI will be separating general mechanisms from those tied to social interactions, as well as characterising the interplay between these mechanisms. This example demonstrates how taking a broader perspective, which includes a strong foothold anchored in diverse literatures, should add value by guiding theoretical expectations and placing important constraints on inferences.

Consequently, the second central pillar of our proposal for HRI research going forward is that knowledge from other cognitive domains, including object perception, memory, attention, and semantics [66–69,71], should be considered together with social cognition research findings. The balance between the role played by these different systems may vary as a function of robot and research context, and it is likely to evolve as experience and expectations change. Therefore, mapping the relationship between these systems across different robots, as well as

over repeated exposures, or after watching the latest Hollywood blockbuster film featuring a robotic protagonist or villain, becomes an important new avenue to probe the structure of associated cognitive and brain systems. In short, we suggest that the default expectation should be that HRI will involve interplay between diverse forms of cognition and experience. Although it currently seems impossible to account for all cognitive systems and processes involved, it is also not sufficient to focus exclusively on social cognition (Figure 2A). Even so, it is clear that the adoption of an integrated approach is still limited, given that adding together knowledge gained from studying interactions with objects, machines, and humans cannot add up to a complete understanding of HRIs (Figure 2B). Therefore, in the following section, we highlight some of the unique challenges facing HRI research, which extend beyond traditional lines of investigation in cognitive and social neuroscience.

Unique Challenges and Opportunities for Studying Human and Social Machines

The third central pillar of the current proposal is that the unique demands presented by interacting with social machines must be better incorporated into models of cognitive and brain function. In the following, we consider salient examples of the challenges and opportunities faced by HRI research. Although not an exhaustive summary, we use these examples to highlight that a central challenge facing cognitive neuroscience research into HRIs is driven by a combination of limited experience, varying expectations, and a wide variety of robots.

Knowledge and experience with objects, tools, animals, and humans involve many concrete examples of canonical forms, features, and functions, which provide a rich set of priors that shape perception and interaction with these objects or agents. This is categorically not the case where robots are concerned. Beyond the wide variety of robot models with bespoke mappings between form and function, robots represent a far more novel category for human perceivers and interactors. The novel nature of robots, particularly **embodied robots** that are encountered *in situ*, means that we have considerably less experience interacting with robots than we do with humans, which has a range of consequences.

One consequence of robot novelty is that we are likely to have vague, uncertain, and unrealistic expectations about what these robots can do or how they might behave, which have been shaped and skewed by media exposure (Box 3). Complicating this issue further are discrepancies in human behaviour toward social robots when encountering such robots 'in the wild' (such as in a shopping mall or airport) compared with in a laboratory. It is estimated that approximately 75% of HRI studies are laboratory based [72] and that the majority of these studies use **Wizard of Oz** methods [73], where the experimenter controls the robot's behaviours and responses unbeknownst to the participant. These facts cast considerable ambiguity over the specificity of the claims being made when it comes to the structure of cognitive systems. Indeed, methods used in laboratory research often unrealistically portray the capabilities of robots, as they depict not autonomous, artificially intelligent robots but human-controlled robots more akin to sophisticated puppets. This reality applies to the vast majority of social robotics research [74] and raises important questions regarding whether inferences from these studies are limited to Wizard of Oz–style methods or might be generalised to interactions with autonomous robots. The distinction is important and highlights a major validity concern for the field, given that the overarching goal of most HRI research is to make the latter inference whilst using the former method.

Another challenge is presented by mismatches between robot form and function. As mentioned previously, some of the most human-like robots, including Kondomoroid (pictured in Figure 1A) or Hanson Robotics Sophia, might superficially appear similar to humans (especially in static photographs), but even the briefest interaction will reveal limited social capacity. Contrast this

Box 3. Managing Expectations and Curbing Enthusiasm: Media's Role in Shaping Human Encounters with Artificially Intelligent S Social Machines

Contemporary society is awash with news articles appearing on an almost daily basis, proclaiming breakthroughs in artificial intelligence and robotics technology. Many of these same articles also include dire narratives that ask whether, as humans, our days as workers, companions, carers, and even lovers are numbered (e.g., [101]). Importantly, however, most people's predictions about what living and working alongside robots might look like is based not on firsthand experience with robots but instead on science fiction depictions of robots found in popular media, including books, films, television, and video games [102,103]. Evidence suggests that people's exposure to robots in the media has an outsized influence on how we perceive robots, because, as roboticist Christoph Bartneck [102] put it, '[A]lmost all our knowledge about robots stems from the media … [and] this tension between expectations fuelled from SciFi and the actual abilities of robots can result in negative experiences' (p. 64). This mismatch between science fiction–fuelled expectations and current robot reality shapes our expectations, and thus how we perceive and interact with artificially socially intelligent machines, in at least two ways. First, depictions of technologically sophisticated and highly intelligent machines, such as those depicted in Figure IA and B, suggest that robots will look and move just like us while also possessing superior powers of perception, strength, computation, and so forth. It is unsurprising that people report being terrified of such agents (e.g., [101,103]) and the roles they might take on in society. The reality is that no currently available embodied robot comes close to cinematic depictions of robots portrayed by leading Hollywood actors. Nevertheless, these dystopian visions shape people's expectations about robots they might encounter in social spaces in the near future. A second consequence of media depictions of robots establishing unrealistic expectations comes from the range of helpful, kind, or socially assistive behaviours performed by robots more often found in the categories depicted in Figure IC, D. Here again, the current robot reality will leave people sorely disappointed if they expect the robot assisting with check-in at the airport might show and evoke as much empathy as Wall-E or demonstrate the loyalty portrayed by BB-8. We would argue that no other category of object or agent is as susceptible to influences beyond one's direct experience, and finding ways to account for media-shaped expectations when encountering these agents (beyond filling out a short questionnaire; e.g., [104,105]) will be crucial for building a more complete understanding of human–machine interaction in social spheres.

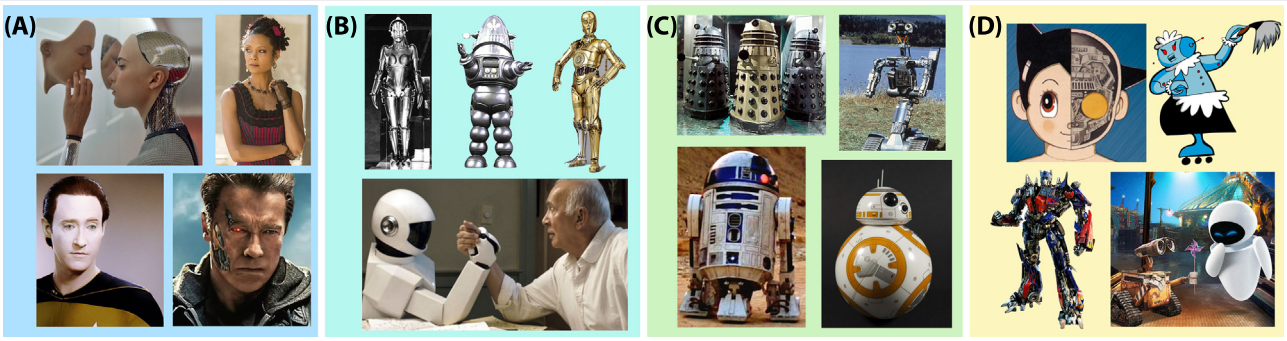

Figure I. Depictions of Artificially Intelligent Social Machines in Popular Media. Such depictions can be roughly categorised into four groups, based on social similarity to real people. (A) Human actors playing the role of highly sophisticated (and often dangerous or deadly) androids, such as (clockwise from top left) Ava in *Ex Machina*, Maeve Millay in *Westworld*, Data in *Star Trek*, and T-800 from *Terminator*. (B) Human actors playing robots but wearing full-body robot suits to disguise identifying human features, including robotic Maria from *Metropolis*, Robby the Robot from *Forbidden Planet*, C-3PO from *Star Wars*, and the robot from *Robot and Frank*. (C) Mechanically animated or human-animated robots that neither look nor move like humans but operate in human spaces to help or harm people, including the Daleks from *Dr. Who*, Number 5 from *Short Circuit*, and R2-D2 and BB-8 from *Star Wars*. (D) Animated depictions of robots unbound by laws of physics and real-world plausibility, including Astro Boy, Rosey the Robot from *The Jetsons*, Optimus Prime, and Wall-E and EVE.

with other artificially intelligent machines, such as Amazon's Echo, which resembles a cylindrical paperweight in appearance, not much larger than a water bottle. Equipped with the Alexa virtual assistant artificial intelligence technology and a human-like voice, however, this seemingly simple object has access to a world of knowledge and can conversationally engage people to a certain extent (though again, it takes little probing for this technology to fail the **Turing Test**). In both instances, the form of the agent is not clearly linked to its expected functional capacity. As a consequence, nonstraightforward links between how form is assessed in perceptual systems and how inferences and expectations regarding functional capacity are generated would need to be built into cognitive and brain models of HRI.

People's general lack of experience with robots in the real world, which can be further exacerbated by cultural and socioeconomic factors (e.g., [29]), also means that the stage of learning about these machines is different from that regarding people and common objects. With such 'immature' representations (in terms of both phylogeny and ontogeny) comes more opportunity for trial-and-error learning rather than consolidated forms of learning. Here again, an earlier

stage of learning means that cognitive and brain systems should be less coherent within and across people, meaning researchers should expect noisier signals at brain and behavioural levels when probing HRIs. Although this has the potential to make interactions between humans and robots less reliable and tractable to study, it also presents an opportunity to study intra- and interindividual differences. Individual differences are clearly at play within the domains of object perception and social cognition (e.g., [75–77]), and we would predict them to be even more pronounced when interacting with social machines due to the vast range of expectations people hold regarding robots in particular.

Because most people are inexperienced in directly interacting with robots, this presents a prime opportunity to explore the impact of experience and learning on the cognitive and brain systems that support HRI. For example, a study examining automatic imitation of actions performed by a human hand and a robotic claw demonstrates that a bias toward human hand actions disappears after repeated exposure to actions performed by a robotic claw [78]. Moreover, when toddlers had access to a small social robot in a childcare setting across a 5-month period, they started to treat this robot more like a peer than a toy [79]. Conversely, other evidence suggests that repeated social interactions with the Cozmo robot (Figure 1A) does not lead to any quantitative or qualitative changes in empathy toward this robot, as measured by behavioural and brain measures [45]. As such, although learning and long-term exposure measures will be vital for exploring human–machine cognition, it will be imperative to carefully design and interpret these studies, keeping in mind that more exposure will not necessarily mean machine- or robot-directed cognition taking on a more social flavour. A better understanding of human brain plasticity during long-term engagement with social machines is vital because the neurocognitive mechanisms that support these interactions will be continually changed by this technology as rapidly as we continue to develop it.

Exploring Future Encounters between Human Minds and Machines

Research using social cognition as a model for examining the mental processes that support interactions with social machines has made valuable contributions to knowledge and will continue to do so [18,19,44,53,54,80]. Building on these initial steps, the central argument of this piece is that we can accelerate progress by broadening the scope of investigation within HRI research to incorporate other domains of cognition that may be equally useful, such as object perception, attention, and many more besides (Figure 2B,C). One key strength of the proposed framework is that it encourages a focus on theory building and highlights the value of posing questions in a broader neurocognitive landscape. A further strength of the framework is that it encourages investigation of how these basic building blocks of cognition flexibly adapt to the widespread variety of social machines, as well as the unusual demands that interacting with such machines may place on cognitive and brain systems. In short, we hope that the proposed framework highlights the inherent difficulty of the task at hand, which is likely to involve a complex and somewhat atypical set of interacting mental pieces. In the following, we consider the implications of the framework for future research, as well as its limitations.

By acknowledging similarities and differences between more established lines of investigation, HRI researchers will be in a better position to set expectations accordingly and respect the inherent challenge of the task at hand (subtext: it's going to be bloody difficult). Conventional psychological and brain science is beset with methodological and theoretical problems that involve the routine use of questionable research practices that produce low levels of reproducibility [81–84]. In the field of social robotics, a growing chorus of both established and emerging researchers is calling for improvements to research quality, including preregistration; replication; and the sharing of data, code, and stimuli [44,74]. On top of this, we have outlined here a number of additional challenges that cognitive (neuro)scientific investigations of HRI present.

At this stage, we would argue that a tone of cautious optimism is most appropriate. The promise and potential of autonomous social machines that serve a number of service and entertainment roles is undoubtedly exciting (as sci-fi has shown us and shaped our expectations for decades; Box 3). In order to reach this goal, however, we must proceed carefully and respect the difficulty of understanding the human side of HRI. Simple experimental manipulations that give inanimate objects human features may have limited explanatory power for developing interactive artificial agents in the real world. Instead, it will be important to consider the multilevel nature of the target, which spans wildly variable expectations, stages of learning, robotic platforms, and degrees of autonomy, amongst others. This is why new theoretical work will be vital for generating and articulating relevant questions as much as testing them [85–87]. Indeed, current challenges cannot be overcome by simply running experiments with better control conditions. Put differently, it would be beneficial to get the questions straight before deploying an army of different robots across different experimental contexts. If not, we will be stuck with the current challenge of how to meaningfully interpret, let alone generalise, findings that emerge. Given the manifold and unique challenges that HRI research faces, providing clarity on scope, relevance, aims, and likely feasibility should be especially key concerns.

Concluding Remarks

In sum, the framework we discuss here aims to harness key insights from cognitive neuroscience to address the novel and unusual demands posed by human interactions with social machines. As a consequence, this framework has an intentionally focused scope, which is unlikely to be directly or immediately relevant to the whole of social robotics. For example, the framework may have less immediate relevance for those with shorter-term applied goals, such as using social robots in clinical or commercial contexts. By explicitly acknowledging such constraints on the generality of our proposal, we hope to help foster a more cumulative science by emphasising what is within and beyond the purview of the proposal [88]. Indeed, as countless researchers have said before, true progress in developing machines that engage humans on a social level will only come from interdisciplinary collaboration across the social, life, and computing sciences (cf. [56]). A more complete understanding of how the human mind and brain negotiate encounters with social machines, therefore, will benefit from engaging the full spectrum of cognitive sciences, including philosophy, psychology, and neuroscience, as theory and empirical findings from these disciplines clearly have much to offer this endeavour (see Outstanding Questions).

Outstanding Questions

Given the unique challenges to researching human interactions with social machines, combined with the reproducibility crisis in psychology and neuroscience in general [80], how should feasibility issues related to establishing the neurocognitive foundations of HRI be addressed?

To what extent are adequately powered replication studies needed to guard against the deleterious and long-lasting impact of initial results that turn out to be false-positive findings?

How much value would be added if HRI hypotheses were more frequently preregistered and specific analyses clearly labelled as confirmatory versus exploratory?

How do general and social cognitive systems interact? How does robot experience shape the influence exerted by these different systems?

What is the relationship (and how does it evolve) between media-shaped expectations about social machines and ongoing firsthand experience?

How can researchers and engineers working to design social machines capitalize upon new insights gained from cognitive science–based investigations into HRI, and how should findings that reveal more complexity and/or murkiness in cognitive systems shape real-life social robotics applications?

Will it ever be possible to design a robot or other type of artificially intelligent machine whose features, appearance, and behaviours align with naive human users' expectations?

How do researchers account for (and even capitalise upon) the continually shifting landscape of our interactions with social machines, given the dynamic nature of brain plasticity, ongoing updates to social machines' software, and myriad ways in which human social intelligence and artificial social intelligence interact?

Acknowledgments

The authors thank #TeamSoBots for helpful feedback on earlier versions of this piece, including Nathan Caruana, Kohinoor Darda, Anna Henschel, Michaela Kent, Ruud Hortensius, and Te-Yi Hsieh. This paper originates from projects in the Social Brain in Action Laboratory that have received funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement number 677270 to E.S.C.), the Leverhulme Trust (PLP-2018-152 to E.S.C.), and Macquarie University (MQRIS: Building Social Robotics Capacity Across Disciplines and Departments; to E.S.C. and R.R.).

References

1. Gates, B. (2008) A robot in every home. *Sci. Am.* 18, 4–11
2. Weinberg, N. (2019) Rbr report: how to start a robotics company. *Robot. Bus. Rev.* 1–12
3. Lockwood, P.L. *et al.* (2020) Is there a 'social' brain? Implementations and algorithms. *Trends Cogn. Sci.* 24, 802–813
4. Caruana, N. and McArthur, G. (2019) The mind minds: the effect of intentional stance on the neural encoding of joint attention. *Cogn. Affect. Behav. Neurosci.* 19, 1479–1491
5. Chao, L.L. *et al.* (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat. Neurosci.* 2, 913–919
6. Chao, L.L. and Martin, A. (2000) Representation of manipulable man-made objects in the dorsal stream. *Neuroimage* 12, 478–484
7. Downing, P.E. *et al.* (2006) Domain specificity in visual cortex. *Cereb. Cortex* 16, 1453–1461
8. Epstein, R. and Kanwisher, N. (1998) A cortical representation of the local visual environment. *Nature* 392, 598–601
9. Kanwisher, N. *et al.* (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 4302–4311
10. Martin, A. *et al.* (1996) Neural correlates of category-specific knowledge. *Nature* 379, 649–652
11. Peelen, M.V. and Downing, P.E. (2005) Within-subject reproducibility of category-specific visual activation with functional MRI. *Hum. Brain Mapp.* 25, 402–408
12. Mahon, B.Z. and Caramazza, A. (2009) Concepts and categories: a cognitive neuropsychological perspective. *Annu. Rev. Psychol.* 60, 27–51

13. Trapp, S. and Bar, M. (2015) Prediction, context, and competition in visual recognition. *Ann. N. Y. Acad. Sci.* 1339, 190–198
14. Adolphs, R. (2009) The social brain: neural basis of social knowledge. *Annu. Rev. Psychol.* 60, 693–716
15. Frith, C.D. and Frith, U. (2012) Mechanisms of social cognition. *Annu. Rev. Psychol.* 63, 287–313
16. Kanwisher, N. (2017) The quest for the FFA and where it led. *J. Neurosci.* 37, 1056–1061
17. Kanwisher, N. (2010) Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc. Natl. Acad. Sci. U. S. A.* 107, 11163–11170
18. Chaminade, T. and Cheng, G. (2009) Social cognitive neuroscience and humanoid robotics. *J. Physiol. Paris* 103, 286–295
19. Cross, E.S. *et al.* (2019) From social brains to social robots: applying neurocognitive insights to human-robot interaction. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 374, 20180024
20. Baraka, K. *et al.* (2020) An extended framework for characterizing social robots. In *Human-Robot Interaction: Evaluation Methods and Their Standardization* (Jost, C. *et al.*, eds), pp. 21–64, Springer
21. Phillips, E. *et al.* (2018) What is human-like? Decomposing robots' human-like appearance using the Anthropomorphic roBOT (ABOT) database. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 105–113, Association for Computing Machinery
22. Sarrica, M. *et al.* (2019) How many facets does a 'social robot' have? A review of scientific and popular definitions online. *Inf. Technol. People* 33, 1–21
23. Kahn, P.H.Ishiguro, H.Friedman, B.Kanda, T. (2006) *What is a Human? - Toward Psychological Benchmarks in the Field of Human-Robot Interaction*, ROMAN 2006 - the 15th IEEE International Symposium on Robot and Human Interactive Communication, Hatfield, pp. 364–371 https://doi.org/10.1109/ROMAN.2006.314461
24. Gray, H.M. *et al.* (2007) Dimensions of mind perception. *Science* 315, 619
25. Gray, K. and Wegner, D.M. (2012) Feeling robots and human zombies: mind perception and the uncanny valley. *Cognition* 125, 125–130
26. Cameron, D. *et al.* (2017) *You Made Him Be Alive: Children's Perceptions of Animacy in a Humanoid Robot*, Springer
27. Cross, E.S. *et al.* (2016) The shaping of social perception by stimulus and knowledge cues to human animacy. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 371, 20150075
28. Klapper, A. *et al.* (2014) The control of automatic imitation based on bottom-up and top-down cues to animacy: insights from brain and behavior. *J. Cogn. Neurosci.* 26, 2503–2513
29. Lim, V.Rooksby, M.Cross, E.S. (2020) Social robots on a global stage: establishing a role for culture during human–robot interaction. *Int. J. Soc. Robot.* https://doi.org/10.1007/s12369-020-00710-4
30. Brown, T.A. and Barlow, D.H. (2009) A proposal for a dimensional classification system based on the shared features of the DSM-IV anxiety and mood disorders: implications for assessment and treatment. *Psychol. Assess.* 21, 256–271
31. Conway, C.C. *et al.* (2019) A Hierarchical Taxonomy of Psychopathology Can Transform Mental Health Research. *Perspect. Psychol. Sci.* 14, 419–436
32. Cuthbert, B.N. and Insel, T.R. (2013) Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* 11, 126
33. Kotov, R. *et al.* (2015) The structure and short-term stability of the emotional disorders: a dimensional approach. *Psychol. Med.* 45, 1687–1698
34. Hu, Y. *et al.* (2018) First Impressions of Personality Traits From Body Shapes. *Psychol. Sci.* 29, 1969–1983
35. Oosterhof, N.N. and Todorov, A. (2008) The functional basis of face evaluation. *Proc. Natl. Acad. Sci. U. S. A.* 105, 11087–11092
36. Over, H. and Cook, R. (2018) Where do spontaneous first impressions of faces come from? *Cognition* 170, 190–200
37. Valentine, T. (1991) A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q. J. Exp. Psychol. A* 43, 161–204
38. Cross, E.S. *et al.* (2012) Robotic movement preferentially engages the action observation network. *Hum. Brain Mapp.* 33, 2238–2254
39. Heider, F. and Simmel, M. (1944) An experimental study of apparent behaviour. *Am. J. Psychol.* 57, 243–259
40. Press, C. *et al.* (2005) Robotic movement elicits automatic imitation. *Brain Res. Cogn. Brain Res.* 25, 632–640
41. Ramsey, R. and Hamilton, A.F.d.C. (2010) Triangles have goals too: understanding action representation in left aIPS. *Neuropsychologia* 48, 2773–2776
42. Reiser, U. *et al.* (2013) Care-O-Bot(r) 3: vision of a robot butler. In *Your Virtual Butler: Lecture Notes in Computer Science 7407* (Trappel, R., ed.), pp. 97–116, Springer
43. Breazeal, C. *et al.* (2016) Social robotics. In *Springer Handbook of Robotics* (Siciliano, B. and Khatib, O., eds), pp. 1935–1972, Springer
44. Henschel, A. *et al.* (2020) Social cognition in the age of human–robot interaction. *Trends Neurosci.* 43, 373–384
45. Cross, E.S. *et al.* (2019) A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 374, 20180034
46. Gazzola, V. *et al.* (2007) The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage* 35, 1674–1684
47. Gobbini, M.I. *et al.* (2011) Distinct neural systems involved in agency and animacy detection. *J. Cogn. Neurosci.* 23, 1911–1920
48. Kilner, J.M. *et al.* (2003) An interference effect of observed biological movement on action. *Curr. Biol.* 13, 522–525
49. Liepelt, R. and Brass, M. (2010) Top-down modulation of motor priming by belief about animacy. *Exp. Psychol.* 57, 221–227
50. Tai, Y.F. *et al.* (2004) The human premotor cortex is 'mirror' only for biological actions. *Curr. Biol.* 14, 117–120
51. Kupferberg, A. *et al.* (2018) Fronto-parietal coding of goal-directed actions performed by artificial agents. *Hum. Brain Mapp.* 39, 1145–1162
52. Breazeal, C. (2007) Sociable robots. *J. Robot. Soc. Jpn.* 24, 591–593
53. Coradeschi, S. *et al.* (2006) Human-Inspired Robots. *IEEE Intell. Syst.* 21, 74–85
54. Hortensius, R. and Cross, E.S. (2018) From automata to animate beings: the scope and limits of attributing socialness to artificial agents. *Ann. N. Y. Acad. Sci.* 1426, 93–110
55. Ishiguro, H. (2013, September 14) Studies on very human-like robots. In *Presented at the International Conference on Instrumentation, Control, Information Technology and System Integration*, Nagoya University, Aichi, Japan
56. Sciutti, A. *et al.* (2018) Humanizing Human-Robot Interaction: On the Importance of Mutual Understanding. *IEEE Technol. Soc. Mag.* 37, 22–29
57. Meltzoff, A. (2007) 'Like me': a foundation for social cognition. *Dev. Sci.* 10, 126–134
58. Collins, E.C. (2019) Drawing parallels in human-other interactions: a trans-disciplinary approach to developing human-robot interaction methodologies. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 374, 20180433
59. Michael, J. and D'Ausilio, A. (2015) Domain-specific and domain-general processes in social perception–A complementary approach. *Conscious. Cogn.* 36, 434–437
60. Ramsey, R. and Ward, R. (2020) Putting the non-social into social neuroscience: A role for domain-general priority maps during social interactions. *Perspect. Psychol. Sci.* 15, 1076–1094
61. Spunt, R.P. and Adolphs, R. (2017) A new look at domain specificity: insights from social neuroscience. *Nat. Rev. Neurosci.* 18, 559–567
62. Bar, M. (2004) Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629
63. Livne, T. and Bar, M. (2016) Cortical Integration of Contextual Information across Objects. *J. Cogn. Neurosci.* 28, 948–958
64. Panichello, M.F. *et al.* (2017) Internal valence modulates the speed of object recognition. *Sci. Rep.* 7, 361
65. Heider, F. (1958) *The Psychology of Interpersonal Relations*, John Wiley & Sons
66. Baddeley, A. (2012) Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29
67. Cabeza, R. and Moscovitch, M. (2013) Memory Systems, Processing Modes, and Components: Functional Neuroimaging Evidence. *Perspect. Psychol. Sci.* 8, 49–55

68. Duncan, J. (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn. Sci.* 14, 172–179
69. Petersen, S.E. and Posner, M.I. (2012) The attention system of the human brain: 20 years after. *Annu. Rev. Neurosci.* 35, 73–89
70. Wykowska, A. (2020) Social robots to test flexibility of human social cognition. *Int. J. Soc. Robot.* https://doi.org/10.1007/s12369-020-00674-5
71. Ralph, M.A. *et al.* (2017) The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* 18, 42–55
72. Baxter, P.Kennedy, J.Senft, E.Lemaignan, S.Belpaeme, T. (2016) *From characterising three years of HRI to methodology and reporting recommendations*, 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, pp. 391–398 https://doi.org/10.1109/HRI.2016.7451777
73. Riek, L.D. (2012) Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *J. Hum. Robot Interaction* 1, 119–136
74. Belpaeme, T. (2020) Advice to new human-robot interaction researchers. In *Human-Robot Interaction: Springer Series on Bio- and Neurosystems* (Jost, C., ed.), pp. 355–369, Springer Nature
75. Chua, K.W. and Gauthier, I. (2020) Domain-specific experience determines individual differences in holistic processing. *J. Exp. Psychol. Gen.* 149, 31–41
76. Turbett, K. *et al.* (2019) Individual Differences in Serial Dependence of Facial Identity are Associated with Face Recognition Abilities. *Sci. Rep.* 9, 18020
77. Bukowski, H. *et al.* (2020) When differences matter: rTMS/fMRI reveals how differences in dispositional empathy translate to distinct neural underpinnings of self-other distinction in empathy. *Cortex* 128, 143–161
78. Press, C. *et al.* (2007) Sensorimotor experience enhances automatic imitation of robotic action. *Proc. Biol. Sci.* 274, 2509–2514
79. Tanaka, F. *et al.* (2007) Socialization between toddlers and robots at an early childhood education center. *Proc. Natl. Acad. Sci. U. S. A.* 104, 17954–17958
80. Wykowska, A. *et al.* (2016) Embodied artificial agents for understanding human social cognition. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 371
81. Open Science Collaboration (2015) PSYCHOLOGY: estimating the reproducibility of psychological science. *Science* 349, aac4716
82. Button, K.S. *et al.* (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376
83. Munafò, M.R. *et al.* (2017) A manifesto for reproducible science. *Nat. Hum. Behav.* 1, 0021
84. Vazire, S. (2018) Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspect. Psychol. Sci.* 13, 411–417
85. Muthukrishna, M. and Henrich, J. (2019) A problem in theory. *Nat. Hum. Behav.* 3, 221–229
86. Oberauer, K. and Lewandowsky, S. (2019) Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* 26, 1596–1618
87. Gray, K. (2017) How to Map Theory: Reliable Methods Are Fruitless Without Rigorous Theory. *Perspect. Psychol. Sci.* 12, 731–741
88. Simons, D.J. *et al.* (2017) Constraints on generality (COG): a proposed addition to all empirical papers. *Perspect. Psychol. Sci.* 12, 1123–1128
89. Kording, K.P.Blohm, G.Schrater, P.Kay, K. (2020) Appreciating the variety of goals in computational neuroscience. In *NBDT* (3), pp. 1–12
90. Posner, M.I. (1980) Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25
91. Driver, J. *et al.* (1999) Gaze Perception Triggers Reflexive Visuospatial Orienting. *Vis. Cogn.* 6, 509–540
92. Frischen, A. *et al.* (2007) Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychol. Bull.* 133, 694–724
93. Samson, D. *et al.* (2010) Seeing it their way: evidence for rapid and involuntary computation of what other people see. *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1255–1266
94. Santiesteban, I. *et al.* (2014) Avatars and arrows: implicit mentalizing or domain-general processing? *J. Exp. Psychol. Hum. Percept. Perform.* 40, 929–937
95. Birmingham, E. *et al.* (2009) Get real! Resolving the debate about equivalent social stimuli. *Vis. Cogn.* 17, 904–924
96. Birmingham, E. and Kingstone, A. (2009) Human social attention. *Prog. Brain Res.* 176, 309–320
97. Willemse, C. and Wykowska, A. (2019) In natural interaction with embodied robots, we prefer it when they follow our gaze: a gaze-contingent mobile eyetracking study. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 374, 20180036
98. Redcay, E. and Schilbach, L. (2019) Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nat. Rev. Neurosci.* 20, 495–505
99. Shamay-Tsoory, S.G. and Mendelsohn, A. (2019) Real-Life Neuroscience: An Ecological Approach to Brain and Behavior Research. *Perspect. Psychol. Sci.* 14, 841–859
100. Rouder, J.N. *et al.* (2016) Is There a Free Lunch in Inference? *Top. Cogn. Sci.* 8, 520–547
101. McClure, P.K. (2018) 'You're fired,' says the robot: the rise of automation in the workplace, technophobes, and fears of unemployment. *Soc. Sci. Comput. Rev.* 36, 139–156
102. Bartneck, C. (2013) Robots in the theatre and the media. In *8th International Conference on Design and Semantics of Form and Movement (DeSForM 2013), Wuxi, China*, pp. 64–70
103. Pederson, I. (2016) Home is where the AI heart is [Commentary]. *IEEE Technol. Soc. Mag.* 35, 50–51
104. Nomura, T. *et al.* (2008) Prediction of human behavior in human-robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Trans. Robot.* 24, 442–451
105. Riek, L.D.Adams, A.Robinson, P. (2011) Exposure to cinematic depictions of robots and attitudes towards them. In *Proceedings of International Conference on Human-Robot Interaction, Workshop on Expectations and Intuitive Human-Robot Interaction*