Trends in Neurosciences

CellPress
REVIEWS

Opinion

Social Cognition in the Age of Human–Robot Interaction

Anna Henschel [1,3,@], Ruud Hortensius [1,3,@] and Emily S. Cross [1,2,*,@]

Artificial intelligence advances have led to robots endowed with increasingly sophisticated social abilities. These machines speak to our innate desire to perceive social cues in the environment, as well as the promise of robots enhancing our daily lives. However, a strong mismatch still exists between our expectations and the reality of social robots. We argue that careful delineation of the neurocognitive mechanisms supporting human–robot interaction will enable us to gather insights critical for optimising social encounters between humans and robots. To achieve this, the field must incorporate human neuroscience tools including mobile neuroimaging to explore long-term, embodied human–robot interaction *in situ*. New analytical neuroimaging approaches will enable characterisation of social cognition representations on a finer scale using sensitive and appropriate categorical comparisons (human, animal, tool, or object). The future of social robotics is undeniably exciting, and insights from human neuroscience research will bring us closer to interacting and collaborating with socially sophisticated robots.

Highlights

As robots become increasingly present in human society, considerable gaps remain between expectations for the social roles these robots might play and their actual abilities.

Research examining social cognition when interacting with robots offers a promising avenue for understanding how best to introduce robots to complex social settings, such as in schools, hospitals, and at home.

Thanks to methodological advances in human neuroscience, such as mobile neuroimaging, human–robot interaction research is moving out of the laboratory and into the real world.

Human Neuroscience as the Icebreaker in a Social Robotics Winter

Human–robot interaction (see Glossary) is a young field currently in a phase of unrest. Since the development of KISMET in the MIT Media Lab in the late 1990s, one of the first social robots, significant progress has been made towards engineering robots capable of engaging humans on a social level. Robots that respond to and trigger human emotions not only enable closer human-machine collaboration, but can also spur human users to develop long-term social bonds with these agents. While progress in developing increasingly innovative and socially capable robots has advanced considerably over the past decade or so, some have suggested that the field is approaching a **social robotics winter**. Referencing the period of disillusionment following escalating hype surrounding artificial intelligence [1], the still-limited social repertoire of even the most advanced embodied robots calls into question the proclaimed 'rise of the social robots' [2,3].

With robots failing to deliver on expectations, social interaction has been named one of the ten grand challenges the field of robotics is now facing [4]. To facilitate progress toward this endeavour, the rich literature of cognitive neuroscience offers vital insights into human social behaviour, not only on a surface level, but also relating to underlying functional and biological mechanisms [5–7]. Both human–robot interaction researchers and neuroscientists working with robots converge in their interest in facilitating smooth and successful social encounters between robots and humans. This joint effort should ultimately enable society at large to take advantage of the often-heralded potential of robots to provide economical care, company, and coaching.

In this opinion article, we argue that studying the human brain when we perceive and interact with robots will provide insights for a clearer and deeper understanding of the human side of human–

[1]Institute of Neuroscience and Psychology, University of Glasgow, 62 Hillhead Street, Glasgow, Scotland, G12 8QB, UK
[2]Department of Cognitive Science, Macquarie University, NSW 2109, Australia
[3]These authors contributed equally to this work

*Correspondence: Emily.Cross@glasgow.ac.uk (E.S. Cross).

[@]Twitter: @annahenschel (A. Henschel), @ruudhortensius (R. Hortensius) and @brain_on_dance (E.S. Cross)

robot interaction, and will thus set the stage for a **social robotics** spring. Our focus on the human side of these interactions, including consideration of the constraints of social cognition, serves to highlight what recent advances in human neuroscience, in terms of method and theory, can contribute to fluent human–robot encounters. The focus of the majority of past studies has been the passive perception of other agents. While this work provides a first step towards characterising social interactions, a focus on perception alone neglects the rich, complex, and dynamic nature of behaviours that unfold during social exchanges in the real world. How can social neuroscience further our understanding of not only perception but also of dynamic relationships with robots? These insights should help explain how people view and treat these artificial agents in relation to humans, pets, and other animals, tools and objects. Moreover, answers to these questions will help us to understand and support resulting societal changes in the domain of care, education, ethics, and law. In reflecting on the neurocognitive machinery that supports human–robot interactions, we suggest that focusing on representations of social cognition and how these change during actual and sustained interactions with physically present robots will be important. Moreover, we argue that minimally invasive mobile neuroimaging techniques offer exceptional promise for deepening our understanding of the human side of human-robot interaction. These methods will accelerate human–robot interaction research by incorporating social dimensions into our exchanges with these machines, thus generating crucial insights helpful in meeting the grand challenge of creating truly social robots. After all, roboticists, neuroscientists, and robots will all benefit from an improved understanding of human social cognition in an age of robots [5,7,8].

The Origins of Imaging the Human Brain During Interactions with Robots

Human fascination with creating a mechanical self dates back to antiquity, with writers in ancient Greece and ancient China conjuring humanlike automata to serve as workers and servants [9]. In the past century, the type of automaton that has most captured the human imagination (and research and development investment) is robots, with some contemporary models edging closer to the fictionalised ideals that first appeared centuries ago. Concurrent with advances in robotics technology has been the advent and rapid development of human brain imaging technology. This technology has been vital in developing our understanding of the neurocognitive mechanisms that support social behaviour among humans. More recently, the fields of human–robot interaction and neuroscience have begun to intersect, providing new vistas on social cognition during interactions with social robots, with seminal studies investigating motor resonance, action observation, joint attention, and empathy felt towards robots. These studies showcase the diversity of brain imaging modalities involved and the technical advances evident from early human–robot interaction research, and provide a starting point for neurocognitive perspectives on these interactions.

One initial study in this domain [10] probed the flexibility of the **action observation network (AON)** and reported that the parts of the parietal, premotor, and middle temporal cortices ascribed to this network respond both to watching humans grasp and manipulate objects, as well as an industrial robot arm performing these same actions. These findings were corroborated by an electroencephalography (EEG) study showing mu-suppression over sensorimotor or AON regions for both robotic and human agents [11]. Insights into motor resonance for robotic actions were further replicated and extended when researchers [12] reported a series of two functional magnetic resonance imaging (fMRI) experiments that found the AON to be, in fact, more strongly engaged during observation of (unfamiliar) robotlike motion, regardless of whether a human or robotic agent performed the movement. These and other initially surprising findings (reviewed in [13]) have been attributed to greater modulation of the AON following greater **prediction errors** due to the unfamiliarity of robotic motion.

Glossary

Action observation network (AON): a collection of brain regions comprising parts of parietal, premotor, and occipitotemporal cortices that responds when watching other agents (human or robotic) in action.
Automatic imitation: see **motor interference.**
Gaze cueing paradigm: a commonly used psychological paradigm used to investigate the mechanisms of joint attention. The gaze of an observed other (human or non-human, physically present or viewed on a screen) either looks towards or away from a visual target the participant is required to attend to, and the cost in a participant's response time is thought to be a measure of social engagement.
Human–robot interaction: see **social robotics.**
Mentalising: a cognitive process by which an individual reflects on, explores, and interprets their own and others' thoughts and feelings, and how these influence behaviour and actions.
Motor interference: observing others perform movements incongruent to one's own has been found to produce motor interference. Motor interference is closely related to automatic imitation, a phenomenon that describes the tendency of humans to implicitly imitate others' actions and other social cues.
Natural language processing: field of study concerned with the recognition and production of natural language by computers and algorithms.
Pain matrix: collection of brain regions associated with empathy and emotional processing when seeing another individual in pain or distress. Primary nodes of this network include bilateral anterior insular and medial anterior cingulate cortices.
Person perception network (PPN): a collection of brain regions responsive to other individuals, especially their faces and bodies. Regions include the fusiform face area and extrastriate body area, among others.
Prediction error: a mismatch between a predicted and observed response.
Repetition suppression: in a brain imaging context, this refers to a reduction in a neural response that emerges when a stimulus (or a certain aspect of a stimulus) is repeated more than once. Also referred to as repetition priming.

While observing robotic movements engages action-related brain areas, questions remain regarding the extent to which human observers also ascribe emotions and intentions to lifeless machines. Past brain imaging studies reveal that humans do indeed show engagement of the **person perception network (PPN)** when observing emotional expressions as expressed by robots [14] and interactions between robots and other humans [15]. The circumstances under which similar brain responses linked to empathy might emerge when observing humans and robots in simulated pain [16,17], or when attempting to decipher the intentions of robots [5], remain an active field of inquiry. An fMRI experiment using the **gaze cueing paradigm** showed behavioural and brain responses linked to **mentalising,** such as enhanced activation of bilateral anterior temporoparietal junction, only when people believed that another person controlled the robot [18].

State-of-the-art Human Neuroscience Approaches to Human–Robot Interaction

Major strides have been made in applying advances in human neuroimaging technology to studying human–robot interactions in contexts that approximate more naturalistic social interactions. These studies further illuminate not only the flexibility and limits of human social cognition when perceiving and interacting with robots, but also some of the challenges and opportunities that roboticists face (and will continue to face) as they develop increasingly social robots. Work in this domain highlights the importance of not only stimulus cues to socialness (i.e., does the agent look and move like a human or a machine?), but also, and arguably even more importantly, how perceivers' prior beliefs or expectations shape brain responses and behaviour [19–21].

Neuroscientists are now also taking advantage of increasingly sophisticated and multivariate analytical approaches to more sensitively probe how the human brain represents robots compared to people (Box 1). Recent work has applied representational similarity analyses to fMRI data collected when participants viewed three agents (a human, an android, and a mechanical-looking robot) performing different actions [22]. Results revealed that different nodes of the AON represent distinct aspects of these actions, and these representations appear to be hierarchically arranged. Specifically, occipitotemporal regions coded for low level action features (such as form and motion integration), while parietal regions coded more abstract and semantic content, such as the action category and intention. These findings corroborate related work that examined effective connectivity between these two nodes when participants viewed actions of varying familiarity [23].

Additional work highlights important aspects of how the human brain computes and evaluates anthropomorphism [24–26]. One study has attempted to evaluate the **uncanny valley hypothesis** using an elegant combination of modelling behavioural ratings and functional connectivity data [25]. The authors reported a response profile within the ventromedial prefrontal cortex that closely reflected the hypothesised, nonlinear, uncanny valley shape when viewing images of robots and humans rated more or less unsettling. Further modelling demonstrated that a distinct signal originating in the amygdala predicted when participants would reject artificial agents. This finding ties in with another recent study [26] that examined anthropomorphising behaviour among a small group of individuals with rare basolateral amygdala lesions. These individuals were able to anthropomorphise animate and living entities similarly to neurologically intact individuals, but anthropomorphised inanimate stimuli (such as a robot) less than controls. The authors suggest that the limbic system plays a key role in processing signals originating from artificial agents in a social versus non-social manner.

However, mere observation of robots in one-off laboratory studies can tell us only so much about human–robot interactions. Two recent fMRI studies highlight further innovations in bringing

Social robotics: this term encompasses a wide variety of research relating to robots designed to engage humans on a social level, often framed in a companionship or assistance context. Human–robot interaction**:** is one facet of this diverse field, which specifically investigates how humans perceive and interact with robots.
Social robotics winter: a term used to describe the current disillusionment surrounding social robots, as technological developments have failed to live up to the hopes and expectations fed by robotic depictions in film, television, and other media, as well as the failure of several recent robotics start-ups.
Theory of mind: the ability to attribute other mental states (thoughts, desires, and intentions) to other individuals. Commonly associated with a network of brain regions.
Theory of mind network: includes the medial prefrontal cortex, bilateral temporoparietal junction, and the precuneus.
Uncanny valley hypothesis: humans prefer anthropomorphic agents, but reject them if they appear too humanlike. To what extent the uncanny valley is an artefact of contemporary experimental procedures remains unknown.
Wizard-of-Oz: describes an experimental set-up in which the robot does not operate autonomously, but rather is controlled by the experimenter, thus resembling the trickster turned wizard in the eponymous film.

Box 1. Delineating the Neural Mechanisms of Human–Robot Interaction

How can we examine the functional and temporal changes in neural representations of social cognition during human–robot interaction? Neuroimaging techniques such as EEG and fMRI provide detailed temporal and spatial information on these changes. Traditionally, researchers have looked at relative differences in measures of neural activity during the perception of human and robotic agents. Most research used univariate analyses thereby focussing on distinct networks in the brain, such as the AON, PPN, and theory-of-mind network. This approach allows researchers to answer questions such as whether brain activation when observing a 'happy' robot is higher or lower compared with observing a happy human. In recent years, however, the development and employment of increasingly more detailed analyses, ranging from repetition suppression, to representational similarity analysis, to multivoxel pattern analysis, provide further and new ways to address questions regarding the overlap of neural architectures for social engagement with humans compared with robots. Repetition suppression enables mapping of potential overlap between similar or dissimilar categories, as repeated stimuli lead to deactivation of regions responsive to these stimuli. For example, does a 'happy' robot followed by a happy human (or vice versa) lead to reduced neural activity in a particular region of interest? The presence of repetition suppression would argue for shared neural resources underlying the processing of perceived robotic and human happiness. The critical next step is to capture the changes in the representation of social cognition during perception and interaction with social robots is the use of multivariate analyses. Representational similarity analyses can establish the similarity in neural activation during the observation of a happy or angry human and a happy- or angry-appearing robot (Figure IA). This approach can test if the neural activation represents a particular stimulus dimension. For example, does activity reflect a representation at the level of agent (activity for robots is dissimilar to humans, regardless of expression) or emotion (activity is dissimilar between happy and angry expressions, but similar across humans or robots). Lastly, a promising way to probe the extent to which perceiving and interacting with humans and robots truly share representations at the neural level is to use multivoxel pattern analyses (Figure IB). Instead of measuring magnitude changes, this technique assesses patterns of neural activity that are predictive of specific task conditions, that is, the representation of different emotions. One way to test possible shared representations is to train a classifier to distinguish the observation of a robot displaying happiness from a robot displaying anger, and to test this classifier to distinguish a human experiencing happiness from experiencing anger. If the human brain represents perceived human and robot emotions similarly, then the decision criteria of the classifier can be used to distinguish these two different categories. Together, these analytical tools provide new vistas on human social cognition during real and long-term interactions with social robots and the representation thereof.

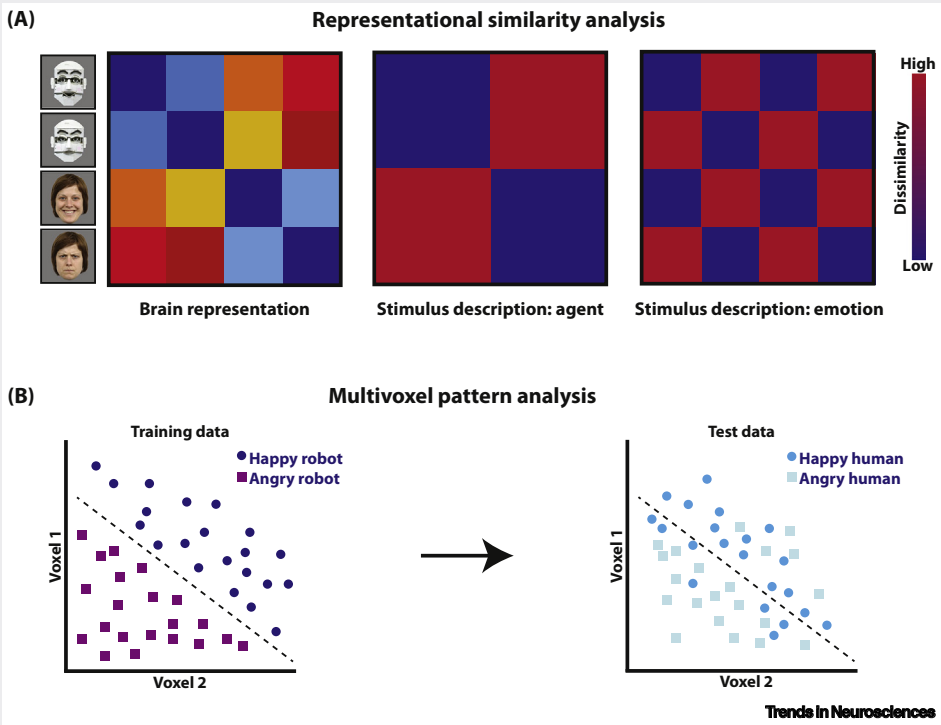

Figure I. Towards a Shared Representation of Social Cognition During Human–Robot Interaction.

together neuroscience, robots, and real-world interactions to advance the fields of social cognition and social robotics collectively. The first study paves the way for future social neuroscience research to incorporate unrestricted social interactions with autonomous agents while simultaneously measuring brain responses [27]. The authors describe a framework that allows participants to interact with a conversational agent (a Furhat robot) or a human partner while a multimodal dataset is collected including behaviour (e.g., speech, eye gaze) and physiology (e.g., respiration, neural activity). Initial results show less engagement of specific brain regions playing a role in everyday social cognition, such as the temporoparietal junction and medial prefrontal cortex, during live human–robot interaction compared with human–human interaction [27]. Another study examined the extent to which a prolonged period of time spent socialising with Cozmo, a palm-sized, playful robot, shapes empathic responses to seeing that same robot 'in pain' [28]. These authors employed pre- and post-socialisation intervention fMRI sessions and measured **repetition suppression** within the **pain matrix** to determine whether a week of daily interactions with Cozmo would shift participants' empathy toward the robot to look more like empathy for another person, based on neural activity as well as behavioural responses. While this study did not find compelling evidence that a week of socialising with this particular robot discernibly shifted empathic responses to look more humanlike [28], this work nonetheless sets the stage for studying the impact of longer-term interactions with robots on social neurocognitive processes. This area of work is crucial if robots will indeed be taking on sustained social roles in close proximity to humans in our daily lives, and should inform robotics developers on ways to maximise social engagement not just for an hour or during an initial encounter, but over the long term.

Together, the findings currently emerging from neuroscientific investigations into human–robot interactions highlight how robots are useful tools for probing core features (actions, emotions, intentions) as well as the flexibility of social cognitive processing in the human brain. While significant progress has been made, efforts to capture and characterise brain responses during live, ongoing interactions with robots remain in the very early stages. As mentioned later, this is likely to be one of the most fruitful areas for further exploration and development. However, before moving forward with real social interactions, clarification is required regarding the engagement of social cognitive brain regions.

How Should We Probe the Neurocognitive Reality of Human–Robot Interaction?

Neural responses, as measured using fMRI and EEG, when perceiving or interacting with robots differ vastly across different brain networks. Generally, activity within the PPN is not reduced when people observe social robots and other artificial agents compared with people, while activity within the **theory-of-mind network** is reduced [5,14]. Going beyond differences in neural activation magnitude, future research in this area will be propelled by mapping the neural representation of social cognition when we engage with robots and characterising how these representations change over time (Box 1).

Many studies examining how humans perceive and interact with robots have focused on the theory-of-mind network and the PPN. These two networks underlie everyday social cognition and are a suitable starting point to investigate the engagement of social cognitive brain regions when encountering robots. Yet, emerging evidence suggests that other brain regions, including the inferior parietal lobule, play a key role when we engage with social robots (Figure 1). Increased activity in object-selective brain regions has consistently been reported across studies using different robotic agents [12,27,28]. It is therefore critical to capture changes beyond the standard person perception and theory-of-mind networks to provide an unbiased account of human–robot interaction, while simultaneously acknowledging the possibility that robots are perceived as objects after all, at least in some respect or in certain circumstances.

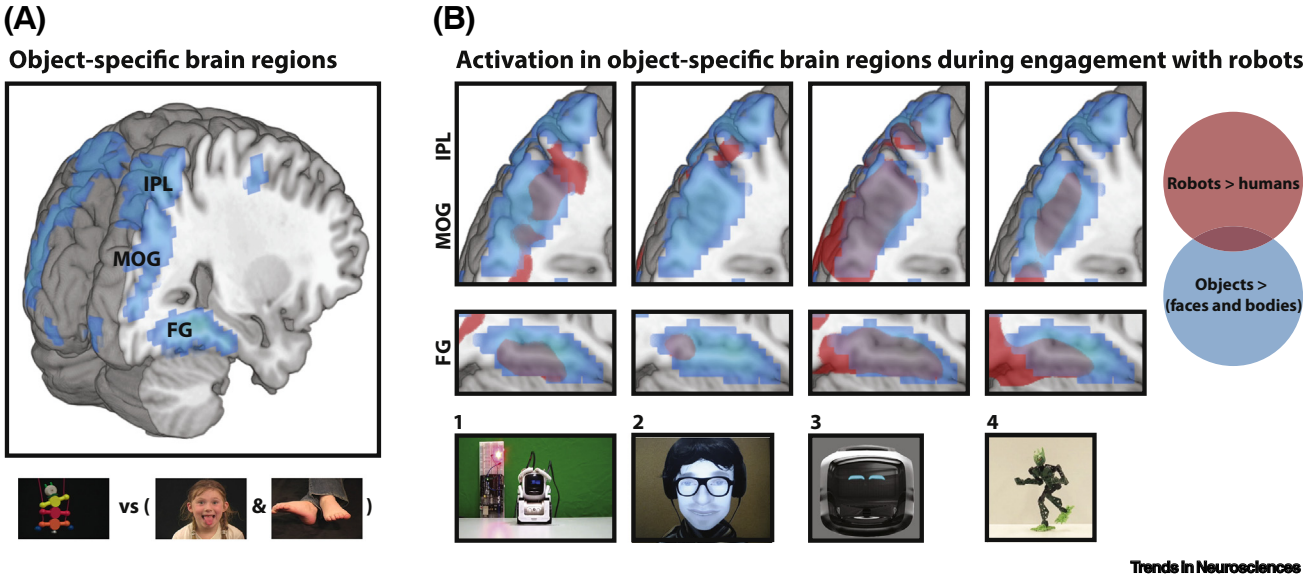

Figure 1. Activity in Object-Specific Brain Regions During Human–Robot interactions. Across several studies that employed different robotic platforms and experimental procedures, a consistent finding is that engaging with robots, compared with engaging with humans, robustly activates object-specific brain regions. (1) Observing robots compared with humans ostensibly experiencing pain or pleasure elicited more activity in the fusiform gyrus (FG), middle occipital gyrus (MOG), and the inferior parietal lobule (IPL) [28]. While (2) live interactions with a robot elicited some of these regions [27], observations of (3) emotions and intentions expressed by a robot (Hortensius and Cross, unpublished data), and (4) robotic movements [12] lead to widespread activity across these regions. These results indicate the importance of considering brain regions that are selective for object perception. Maps for each study are overlaid on top of an independent object localiser [67]. Unthresholded group-maps are shown for the four studies, while the objects versus faces and bodies statistical map (n = 28) for the object localiser is visualised at the family-wise error (FWE) corrected threshold of P <0.05 (k = 10). Data for (1) and (2) are from https://identifiers.org/neurovault.image:108836 and https://identifiers.org/neurovault.image:112530 respectively [68]. Abbreviations: FG, fusiform gyrus; MOG, middle occipital gyrus; IPL, inferior parietal lobule.

Researchers have almost exclusively tested whether robots elicit humanlike responses (i.e., do we perceive and react to emotions expressed by a robot similarly to those expressed by a human?). Focusing on direct comparisons between robots and humans does not acknowledge the possibility that robots could elicit sub- or supra-threshold brain responses in relation to a particular object category. Increased activity in response to human stimuli could therefore be the result of a narrow (univariate) comparison between the two agent categories. A central question in human–robot interaction studies should be what the appropriate comparison categories are for different types of robots. Of course, these could range from humans to objects to animals, and the best answer will naturally depend on the specific research question being tested [29]. To establish the place robots might occupy in our social milieu, we need to measure the (dis)similarity to animate agents (e.g., a human or pet) as well as objects (e.g., a phone). Answers to these questions will not only advance our understanding of how people perceive robots and the development of psychological benchmarks for the success of social robots, but also touch upon philosophy, cognitive science and law, which have important implications for society at large (e.g., morality and ethics) [30–32].

Towards Understanding Real Interactions with Social Robots

Screen-based experiments, third-person observation, and one-off or short-term interactions with robots already provide crucial insights on the social cognitive processes that underlie engagement with these novel agents. For the field to move forward, future studies should investigate real and long-term interactions with embodied robots in ecologically valid settings. These studies will provide much needed evidence as to how the human brain negotiates interactions with these agents in the real world. Interactions in social spaces that go beyond the laboratory and are

relevant to the robotic platform and the user (e.g., schools, care facilities, and hospitals) will be particularly important [9]. The field of social robotics has a long tradition of usability and user experience studies and these investigations will benefit from the sharpened focus on rigor and reproducibility that contemporary psychology and neuroscience bring to the table (Box 2).

The field of social neuroscience in general still needs to answer the call for taking into account the importance of the second person in an interaction [33]; this challenge is especially relevant for the study of our interactions with social robots. Paradigms employing free-flow interactions, wherein a recursive perception-action loop exists between two or more agents, are needed. Fortunately, several studies have begun to look at the impact of exposure to (or interactions with) robots, which cover a wider variety of robot design and morphology [8,18]. This work is starting to explore neurocognitive aspects of human–robot interactions by integrating information derived from behaviour (e.g., speech and eye gaze) and physiology (e.g., respiration and neural activity) [27,28]. One of the next steps towards measuring truly unrestricted social interactions is through the use of mobile functional near-infrared spectroscopy (as highlighted later). Combining these state-of-the-art neuroscience methods with new developments in **natural language processing** should enable researchers to step away from **Wizard-of-Oz methods** and provide new ways to examine the social nature of human–robot interactions.

Human–robot interactions are shaped by prior experiences, expectations, and beliefs that are continuously updated [5]. It is therefore critical to go beyond contrasting pre- versus post-interaction measures, and incorporate longitudinal experimental designs to address questions on experience-dependent plasticity of human social cognition when interacting with social robots. Of note, several commercially available robots allow researchers to collect large datasets per experimental subject over long periods of time, somewhat akin to the experience sampling method (an intensive longitudinal collection of self-report measures). For example, the Cozmo robot [28,34] collects a rich set of data spanning facial recognition, game performance, and

Box 2. Integrating Open Science Practices into Human–Robot Interaction Studies

The movement towards open science practices and an increased focus on the reproducibility of research findings is gaining momentum across research domains in the life and physical sciences, including psychology and human neuroscience [56,57]. Similarly, these issues are acknowledged in artificial intelligence (AI) research [58], and have recently been further reflected upon by robotics researchers [59–61]. Issues of transparency and reproducibility are especially important for investigations of the neurocognitive mechanisms supporting human–robot interaction. Integrating methods and tools from psychology and neuroscience, researchers not only face reproducibility issues key to these fields (e.g., reliability of fMRI findings [62], and researchers' degrees of freedom in preprocessing pipelines of fNIRS and fMRI data [63,64]), but also issues specific to the field of social robotics (e.g., cross-platform generalisability and access to expensive and bespoke robotic platforms). Encouragingly, experimental reform is being implemented in the human–robot interaction community, with the 2020 ACM/ IEEE International Conference on Human–Robot Interaction being the first to invite replication studies for submission. In recent years, psychologists and neuroscientists are more broadly embracing open science practices, which will help to remedy many of the above-mentioned issues. Concrete actions along these lines include taking steps like preregistering studies, conducting replication studies, sharing research materials, and (anonymized) data, as well as posting preprint articles [56,57]. This scientific reform can especially benefit human–robot interaction research, as studies are often resource- and time-intensive and include relatively small samples of subjects. Sharing data and scripts will enable the wider community to conduct secondary and meta-analyses and exploratory tests on published data. Sharing of research resources and products should also contribute to a more inclusive community, giving, for example, access to data from bespoke robotic platforms. Finally, a movement toward greater openness and transparency should facilitate more exchange between disciplines as well as a more robust human–robot interaction literature, by creating an ecosystem conducive of cross-platform replication. One question the field needs to address is the cross-platform generalisability of previous findings [14,65]. Developmental social robotics already successfully implements artificial architectures to test cross-platform generalisability [66], and future research should further incorporate this practice to replicate and extend previous findings. Moving forward, the implementation of open science practices can help facilitate more reproducible (and thus reliable) user studies, and can foster a common ground in terms of methodology between human–robot interaction researchers and cognitive neuroscientists.

'emotional responses' performed by the robot. Of course, these procedures must consider privacy, data protection, and other ethical issues [35], but nonetheless offer promise if employed responsibly.

A consideration to keep in mind in the context of social cognition when interacting with robots is the target population that the robots are designed for, and the purpose of these interactions. Whereas two key target populations for social robotics are children and older adults, participant samples in neuroscience and psychology predominately comprise young adults, and are often biased towards specific sectors of society (e.g., educated and a relatively high socio-economic status) [36]. Further, cultural variation exists in the acceptance and uptake of robots [9], and this cultural heterogeneity is not fully represented in basic research, which tends to be conducted in industrialised countries, often in western ones. As research on human–robot interaction gradually moves towards broader geographical and societal representation, it is important to consider differences in expectations, attitudes, and beliefs, as well as in prior experiences with robots. This variation needs to be considered in the forms of individual differences (e.g., in learning and plasticity), as well as differences between age groups (e.g., [37]) and cultures. For example, one needs to take into consideration that countries such as Japan and South Korea have a longer tradition of research and development in this area than most western countries [38–40]. Similar to an individualised approach that many technology companies adopt (e.g., social media and streaming services), for which cognitive neuroscience has also advocated [41], the time is ripe for research into human–robot interaction to adopt methods that are sensitive to and capitalise upon individual differences. Considering how quickly people adopt and can adapt to new technologies, as well as the impact of potential generational differences on attitudes towards such technologies, and the continuous development of new social robotics platforms, it is imperative to keep in mind what a fast-moving and continuously evolving target human–robot interaction is. In order for research in this dynamic area to maximise relevance and generalisability, specialised methods that enable researchers to map this variation are required. Combining real and extended interactions with continuous data collection, neuroscience methods and machine learning, could thus be a major step towards personalised human–robot interaction [42].

The Promises and Pitfalls of Using Mobile Brain Imaging in Embodied Human–Robot Interaction Studies

New developments in mobile neuroimaging techniques provide the necessary testing ground for how robots might resonate at the social level. One promising technique for studying human–robot interactions is functional near infrared spectroscopy (fNIRS). This technique has been advancing steadily since a connection between human brain function and corresponding light absorption was originally established [43]. This imaging modality, like fMRI, maps the blood oxygen level dependent response, taking advantage of the transparency of biological tissue (such as skin and bone) in the near-infrared spectrum (for a comprehensive review see [44]). Light shone on the head with laser diodes or LEDs travels through the skull, scatters back in a banana-shaped curve and is eventually picked up by a detector located at approximately 3 cm separation. The constraints of fNIRS relate to its relatively shallow penetration depth (reaching the outer layers of the cerebral cortex) and relatively low spatial (2–3 cm) and temporal resolution (up to 10 Hz). It has a lower spatial resolution than fMRI and a slower temporal resolution than EEG, yet brings the advantages of being cost effective, portable and relatively robust to movement artefacts.

These advantages allow for mobile and unobtrusive neuroimaging, thus presenting fNIRS as an optimal candidate for conducting embodied human–robot experiments, especially with under-represented groups such as young children, patients, and older adults that often cannot

participate in more constraining types of data collection. Researchers in human–robot interaction have embraced fNIRS as a tool to construct feedback loops to control robotic movement or behaviour, [45] and as an implicit response evaluation to various robotic systems ([46–48], for a review see [49]). Various high-quality commercial imaging systems that allow high-density channel and hyper-scanning set-ups (with great potential for research on dyads or groups interacting with a robot) are now available and a recent proof of concept study shows the possibility of using fNIRS for connectivity analyses [50].

The transition from laboratory-constrained experiments that employ screen-based evaluations of social robots, to the measurement of unrestricted real-world interactions with physically embodied robots using fNIRS, should be a gradual process, adding complexity in a stepwise fashion (Figure 2). For example, in recent years, the brain networks involved in observing social interaction have been mapped in detail [51]. Two regions, the posterior superior temporal sulcus (STS) and the temporoparietal junction (TPJ), code different aspects of observed interactions [52–54]. A logical next question is the extent to which the presence and content of interactions with robots is also coded in these regions in third-person encounters. Following on, insights gained from these experiments will pave the way for an embodied research approach where brain activity can be measured during real interactions between humans and robots in

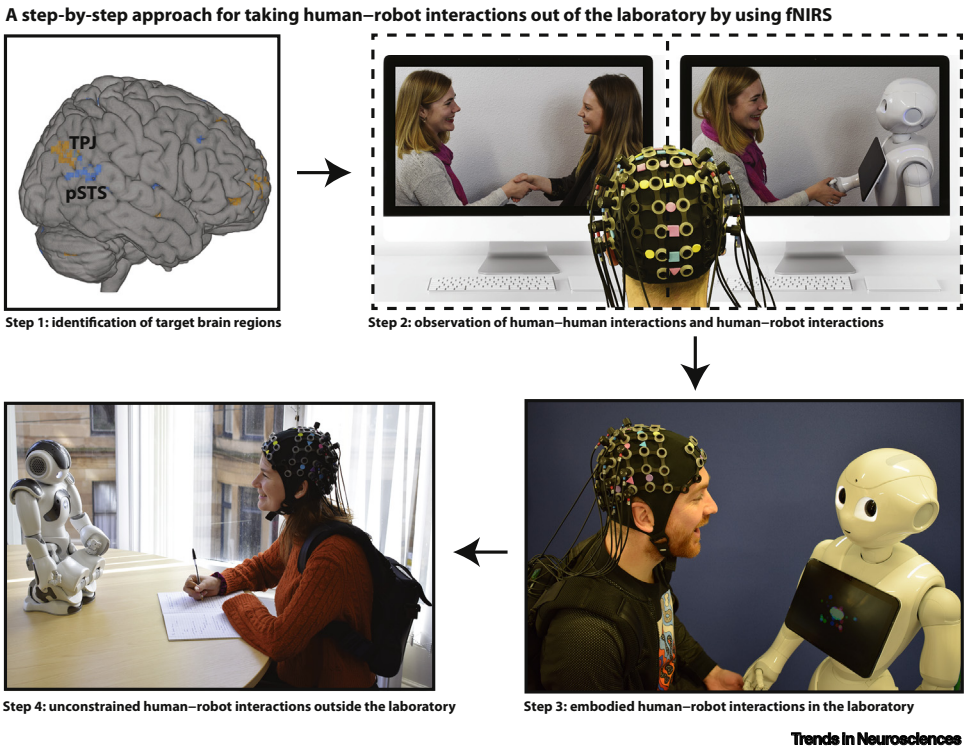

Figure 2. Employing Functional Near-Infrared Spectroscopy for Unconstrained Human–Robot Interactions. A stepwise approach can be undertaken to allow for unconstrained human–robot interactions outside the laboratory in the real world. A first step is the identification of brain regions implicated in a social cognitive process of interest as identified in previous findings (e.g., literature and pilot studies). This is followed by a screen-based exploration of the involvement of these regions during the observation of human–human and human–robot interactions. A third step is the relatively unconstrained interaction with a robot in the context of a laboratory, followed by a final step that allows for embodied interactions with a robot in everyday environments (e.g., schools and homes). The result of each step can inform the methodology and analysis employed in the next step. Photographs provided by Michaela Kent, Anna Henschel, and Rebecca Smith.

unconstrained interactions. In a recent study, for instance, the authors used a general linear model (GLM)-based analysis to automatically identify functional events in fNIRS data, and employed a 'brain-first' approach, where instead of being constrained by a block- or event-related task design, a more ecologically valid setting can be chosen [55]. One can envision applying similar methodologies in the context of human–robot interaction experiments.

When using fNIRS in embodied interaction experiments with social robots, several decisions need to be taken: (i) will the device be used to control the robot or inform the evaluation of the robot?; (ii) how long and 'natural' or unconstrained can the interaction be and still yield reliable and interpretable data? Most fNIRS systems, while lightweight and portable in a fitted backpack, cannot be worn for longer than about 45 min, due to the pressure of the optodes on participants' scalps. When performing games or tasks that involve joint movement, another important limitation to keep in mind is that most commercially available social robots are not capable of repeating the same motions for hours on end, as motors can overheat and batteries run out. However, despite these constraints, using fNIRS in embodied social robotics studies promises to take us one step closer to following the tenets of a two-person neuroscience [33]. Only by freeing the robots from the screen can we begin to understand how embodied interactions affect cognitive processes in socially relevant areas of the cortex, including the superior temporal sulcus, temporoparietal cortex and frontal cortex.

Concluding Remarks

Neuroscience-informed human–robot interaction is making important advances in changing the landscape of social robotics, while concurrently deepening our understanding of the human brain. Beyond perceiving robots in screen-based experiments, recent insights have shown that more sophisticated analysis methods and the trend of gathering data during real-time, embodied interactions with robots, can deepen our knowledge of core mechanisms supporting social cognition. An added (and natural) benefit to this basic human neuroscience research, is that it also stands to inform the development and design of next generation social robots, the same robots that may eventually become social companions that provide support and care. That being said, just over a decade of neuroscientific contributions to human–robot interaction have shown that major questions still remain, for instance: (i) how does the sophisticated neural machinery of the human brain support our interactions with these novel, mechanical companions?; (ii) how does the representation of social cognition change over time as robots become more deeply integrated into our social life (see Outstanding Questions)? Insights from future studies combining human neuroscience and social robotics will prepare us for a future of living with autonomous robots that resonate with us at the social level.

Acknowledgements

The authors thank Michaela Kent, Te-Yi Hsieh, Laura Jastrzab, Ben Jones, Richard Ramsey, and Kohinoor Darda for comments on earlier versions of this piece. This paper originates from projects in the Social Brain in Action Lab that have received funding from the European Research Council (ERC) 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 the Bial Foundation (to R.H.).

References

1. Natale, S. and Ballatore, A. (2020) Imagining the thinking machine. *Convergence Int. J. Res. New Media Technol.* 26, 3–18
2. Tulli, S. *et al.* (2019) Great expectations & aborted business initiatives: The paradox of social robot between research and industry. *CEUR Workshop Proceedings* 2491, 1–10
3. Campa, R. (2016) The rise of social robots: A review of the recent literature. *J. Evol. Technol.* 26, 106–113
4. Yang, G.Z. *et al.* (2018) The grand challenges of science robotics. *Sci. Robot.* 3
5. Hortensius, R. and Cross, E.S. (2018) From automata to animate beings: the scope and limits of attributing socialness to artificial agents: Socialness attribution and artificial agents. *Ann. N. Y. Acad. Sci.* 1426, 93–110
6. Agnieszka, Wykowska *et al.* (2016) Embodied artificial agents for understanding human social cognition. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20150375
7. Chaminade, T. and Cheng, G. (2009) Social cognitive neuroscience and humanoid robotics. *J. Physiol. Paris* 103, 286–295

Outstanding Questions

What are the scope and limitations of social cognition when interacting with social robots? Beyond responding to movement, recognising emotions, and incorporating gaze behaviour of the robot into the equation, are we able to feel empathy for, attribute intentions to, and collaborate with these mechanical beings? Can we form meaningful social relationships with them? Will it ever be possible to develop a robot with a range of social cognitive abilities that resembles (or even improves upon) that of humans?

How do long-term interactions with social robots shape social cognition? Could the human brain's representation of emotions expressed by a robot ever become indistinguishable from the representation of emotions expressed by a human? To what extent can neurocognitive processes be repurposed during human–robot interaction, resulting in shared representations of social cognition when humans or robots are involved?

Do robots need to be framed as social agents at all in order to be useful in social contexts? Or are there some situations (e.g., elderly care) where social robots are perhaps most successful and useful when introduced simply as 'tools'? While most studies focussed on testing the extent to which robots elicit responses similar to humans, might it be more instructive to assign robots to their own distinct category, which stands apart from the categories of animate agents (e.g., a human or pet) and objects (e.g., a phone)?

Establishing the neural mechanisms supporting human–robot interaction beyond the theory-of-mind network and PPN, what role do object-specific brain regions play during human–robot interaction?

With the field moving towards naturalistic interactions, to what extent will previous findings from the laboratory on passive observation of robots (whether *in situ* or on screens) replicate and generalise to the real world? Also, to what extent do findings replicate across robotic classes (e.g., humanoid vs. mechanoid vs. animal-like) and platforms?

8. Wiese, E. *et al.* (2017) Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Front. Psychol.* 8
9. Broadbent, E. (2017) Interactions with robots: The truths we reveal about ourselves. *Annu. Rev. Psychol.* 68, 627–652
10. Gazzola, V. *et al.* (2007) The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage* 35, 1674–1684
11. Oberman, L.M. *et al.* (2007) EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing* 70, 2194–2203
12. Cross, E.S. *et al.* (2012) Robotic movement preferentially engages the action observation network. *Hum. Brain Mapp.* 33, 2238–2254
13. Press, C. (2011) Action observation and robotic agents: Learning and anthropomorphism. *Neurosci. Biobehav. Rev.* 35, 1410–1418
14. Hortensius, R. *et al.* (2018) The perception of emotion in artificial agents. *IEEE Trans. Cogn. Dev. Syst.* Published online April 19, 2018. https://doi.org/10.1109/TCDS.2018.2826921
15. Wang, Y. and Quadflieg, S. (2015) In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions. *Soc. Cogn. Affect. Neurosci.* 10, 1515–1524
16. Suzuki, Y. *et al.* (2015) Measuring empathy for human and robot hand pain using electroencephalography. *Sci. Rep.* 5, 1–9
17. Rosenthal-von der Pütten, A.M. *et al.* (2014) Investigations on empathy towards humans and robots using fMRI. *Comput. Hum. Behav.* 33, 201–212
18. Özdem, C. *et al.* (2017) Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Soc. Neurosci.* 12, 582–593
19. 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
20. Cross, Emily S. *et al.* (2016) The shaping of social perception by stimulus and knowledge cues to human animacy. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20150075
21. Gowen, E. *et al.* (2016) Believe it or not: Moving non-biological stimuli believed to have human origin can be represented as human movement. *Cognition* 146, 431–438
22. Urgen, B.A. *et al.* (2019) Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. *Neuropsychologia* 127, 35–47
23. Gardner, T. *et al.* (2015) Dynamic modulation of the action observation network by movement familiarity. *J. Neurosci.* 35, 1561–1572
24. Wiese, E. *et al.* (2018) Seeing minds in others: Mind perception modulates low-level social-cognitive performance and relates to ventromedial prefrontal structures. *Cogn. Affect. Behav. Neurosci.* 18, 837–856
25. Pütten, A.M.R. der *et al.* (2019) Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *J. Neurosci.* 39, 6555–6570
26. Waytz, A. *et al.* (2018) Anthropomorphizing without social cues requires the basolateral amygdala. *J. Cogn. Neurosci.* 31, 482–496
27. Birgit, Rauchbauer *et al.* (2019) Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20180033
28. Cross, Emily S. *et al.* (2019) A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20180034
29. Collins, Emily C. (2019) Drawing parallels in human–other interactions: a trans-disciplinary approach to developing human–robot interaction methodologies. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20180433
30. Prescott, T.J. (2017) Robots are not just tools. *Connect. Sci.* 29, 142–149
31. Bigman, Y.E. *et al.* (2019) Holding robots responsible: The elements of machine morality. *Trends Cogn. Sci.* 23, 365–368
32. Kahn Jr., P.H. *et al.* (2007) What is a human?: Toward psychological benchmarks in the field of human–robot interaction. *Interact. Stud.* 8, 363–390
33. Schilbach, L. *et al.* (2013) Toward a second-person neuroscience 1. *Behav. Brain Sci.* 36, 393–414
34. Ciardo, F. *et al.* (2020) Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition* 194, 104109
35. Rafaeli, A. *et al.* (2019) Digital traces: New data, resources, and tools for psychological-science research. *Curr. Dir. Psychol. Sci.* 28, 560–566
36. Henrich, J. *et al.* (2010) The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83
37. Kirsch, L.P. *et al.* (2018) Dance training shapes action perception and its neural implementation within the young and older adult brain. *Neural Plast.* 2018, 1–20
38. Hinz, N.-A. *et al.* (2019) Individual differences in attitude toward robots predict behavior in human-robot interaction. In *Social Robotics. ICSR 2019. Lecture Notes in Computer Science* (vol 11876) (Salichs, M. *et al.*, eds), pp. 64–73, Springer, Cham
39. Perez-Osorio, J. *et al.* (2019) More than you expect: Priors influence on the adoption of intentional stance toward humanoid robots. In *Social Robotics. ICSR 2019. Lecture Notes in Computer Science* (vol 11876) (Salichs, M. *et al.*, eds), pp. 119–129, Springer, Cham
40. Cameron, D. *et al.* (2017) You made him be alive: Children's perceptions of animacy in a humanoid robot. In *Biomimetic and Biohybrid Systems. Living Machines 2017. Lecture Notes in Computer Science* (vol 10384), pp. 73–85, Springer, Cham
41. Gordon, E.M. *et al.* (2017) Precision functional mapping of individual human brains. *Neuron* 95, 791–807.e7
42. Clabaugh, C. and Matarić, M. (2018) Robots for the people, by the people: Personalizing human-machine interaction. *Sci. Robot.* 3, eaat7451
43. Chance, B. *et al.* (1993) Cognition-activated low-frequency modulation of light absorption in human brain. *PNAS* 90, 3770–3774
44. Pinti, P. *et al.* (2018) The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Ann. N. Y. Acad. Sci.* 34, 269
45. Solovey, E. *et al.* (2012) *Brainput: Enhancing interactive systems with streaming fnirs brain input*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 2193–2202
46. Strait, M. and Scheutz, M. (2014) *Measuring users' responses to humans, robots, and human-like robots with functional near infrared spectroscopy*, The 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, pp. 1128–1133
47. Kawaguchi, Y. *et al.* (2012) *Investigation of brain activity after interaction with seal robot measured by fNIRS*, Proceedings - IEEE International Workshop on Robot and Human Interactive Communication, Paris, France, pp. 571–576
48. Nuamah, J.K. *et al.* (2019) Neural efficiency of human–robotic feedback modalities under stress differs with gender. *Front. Hum. Neurosci.* 13
49. Canning, C. and Scheutz, M. (2013) Functional near-infrared spectroscopy in human-robot interaction. *J. Human-Robot Interaction* 2, 62–84
50. Bulgarelli, C. *et al.* (2018) Dynamic causal modelling on infant fNIRS data: A validation study on a simultaneously recorded fNIRS-fMRI dataset. *NeuroImage* 175, 413–424
51. Quadflieg, S. and Koldewyn, K. (2017) The neuroscience of people watching: how the human brain makes sense of other people's encounters. *Ann. N. Y. Acad. Sci.* 1396, 166–182
52. Walbrin, J. *et al.* (2018) Neural responses to visually observed social interactions. *Neuropsychologia* 112, 31–39
53. Walbrin, J. and Koldewyn, K. (2019) Dyadic interaction processing in the posterior temporal cortex. *NeuroImage* 198, 296–302
54. Isik, L. *et al.* (2017) Perceiving social interactions in the posterior superior temporal sulcus. *PNAS* 114, E9145–E9152
55. Pinti, P. *et al.* (2017) A novel GLM-based method for the automatic identification of functional events (AIDE) in fNIRS data recorded in naturalistic environments. *NeuroImage* 155, 291–304

What are the individual, cultural, and developmental constraints of human–robot interaction? How best can we incorporate findings from ongoing work examining questions in these domains to create more diverse, adaptable, and engaging robots?

Does the field need a unifying theoretical framework to explain how robots impact different aspects of social cognition (e.g., empathy or reward)?

56. Munafò, M.R. *et al.* (2017) A manifesto for reproducible science. *Nat. Hum. Behav.* 1, 0021
57. Poldrack, R.A. *et al.* (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* 18, 115–126
58. Hutson, M. (2018) Artificial intelligence faces reproducibility crisis. *Science* 359, 725–726
59. Bethel, C.L. and Murphy, R.R. (2010) Review of human studies methods in HRI and recommendations. *Int. J. Soc. Robot.* 2, 347–359
60. Eyssel, F. (2017) An experimental psychological perspective on social robotics. *Robot. Auton. Syst.* 87, 363–371
61. Irfan, B. *et al.* (2018) *Social psychology and human-robot interaction: An uneasy marriage*, Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Chicago, IL, USA, pp. 13–20
62. Button, K.S. *et al.* (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376
63. Pinti, P. *et al.* (2019) Current status and issues regarding preprocessing of fNIRS neuroimaging data: an investigation of diverse signal filtering methods within a general linear model framework. *Front. Hum. Neurosci.* 12
64. Carp, J. (2012) The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage* 63, 289–300
65. Cross, Emily S. *et al.* (2019) From social brains to social robots: applying neurocognitive insights to human–robot interaction. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20180024
66. Cangelosi, A. and Schlesinger, M. (2018) From babies to robots: The contribution of developmental robotics to developmental psychology. *Child Dev. Perspect.* 12, 183–188
67. Pitcher, D. *et al.* (2011) Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage* 56, 2356–2363
68. Gorgolewski, K.J. *et al.* (2015) NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinform.* 9, 3743–3749