PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Opinion piece

Cite this article: Skewes J, Amodio DM, Seibt J. 2019 Social robotics and the modulation of social perception and bias. *Phil. Trans. R. Soc. B* **374**: 20180037.
http://dx.doi.org/10.1098/rstb.2018.0037

Accepted: 11 January 2019

One contribution of 17 to a theme issue 'From social brains to social robots: applying neurocognitive insights to human–robot interaction'.

Subject Areas:
cognition

Keywords:
social, robot, stereotypes, bias, prejudice, cognition

Author for correspondence:
Joshua Skewes
e-mail: filjcs@cas.au.dk

Social robotics and the modulation of social perception and bias

Joshua Skewes[1], David M. Amodio[3,4] and Johanna Seibt[2]

[1]Department for Linguistics, Cognitive Science and Semiotics, and Interacting Minds Center, and [2]Research Unit for Robophilosophy, School of Culture and Society, Aarhus University, Denmark
[3]Department of Psychology and Neural Science, New York University, New York, NY, USA
[4]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

DMA, 0000-0001-7746-0150

The field of social robotics offers an unprecedented opportunity to probe the process of impression formation and the effects of identity-based stereotypes (e.g. about gender or race) on social judgements and interactions. We present the concept of fair proxy communication—a form of robot-mediated communication that proceeds in the absence of potentially biasing identity cues—and describe how this application of social robotics may be used to illuminate implicit bias in social cognition and inform novel interventions to reduce bias. We discuss key questions and challenges for the use of robots in research on the social cognition of bias and offer some practical recommendations. We conclude by discussing boundary conditions of this new form of interaction and by raising some ethical concerns about the inclusion of social robots in psychological research and interventions.

This article is part of the theme issue 'From social brains to social robots: applying neurocognitive insights to human–robot interaction'.

1. Introduction

The term 'social robot' invokes images of prototypical devices such as NAO, I-Cub or Pepper—robotic devices with humanoid shapes, which can elicit smiles, verbal greetings or even hugs from people [1]. However, scientific definitions of the class of robots that qualify as 'social' are not easy to come by, for it is not yet clear how conventional definitions of *sociality* apply to robotic devices or to the human reactions they elicit. Early characterizations of some social robots (cf. [2,3]) fail to address this difficulty. For the present purposes, social robots may be practically defined as programmable devices that are designed to act within the physical and symbolic space of human social interactions, and which have affordances for what people common-sensically call *social interactions*. Indeed, people easily engage with such devices, following typical patterns of social interactions, while experiencing both similarities and distinctive discrepancies from 'authentic' interactions with a human partner. For these reasons, social robots promise to provide a useful set of devices for investigating human social behaviour and cognition, as well as for customizing and designing robot-mediated social interaction [4].

In this article, we illustrate how a new form of robot-mediated social interaction might be used to link research in social robotics to research in social cognition aimed at debiasing effects of social stereotypes on judgements and decisions. In §2, we describe the role of stereotypes in social judgement, providing a context for the potential role of social robotics in debiasing interventions. In §3, we introduce the concept of fair proxy communication (FPC [5,6]). FPC is a novel communicational form whereby robots are used to intervene on human interactions to make them more equitable, by decreasing stereotyping and social bias. We present an abbreviated definition of FPC and briefly describe ongoing research, particularly as it relates to social cognition and bias. In §4, we present and address four major concerns for the FPC approach to social cognition. In §5, we discuss how FPC may be implemented in behavioural and neuroimaging research to illuminate

THE ROYAL SOCIETY PUBLISHING

questions about impression formation and bias in social cognition. Sections 6 and 7 illustrate one direction for research and intervention in further detail, invoking theory and research from social judgement heuristics and describing how a formal model of FPC-based interaction sheds light on its utility as an intervention. After discussing some important caveats, limitations and other considerations concerning FPC in §8, we conclude by noting important theoretical and ethical considerations for future research using FPC systems.

2. Implicit bias and the promise of social robotics

Our first impression of another person emerges from myriad cues, including information about identity (e.g. appearance, gender, social identity), behaviour (e.g. facial expressions, gesture, proximity, voice, even smell) and situational factors [7]. How we interpret another person's behaviour depends not only on our perceptions of their actions, but also on prior assumptions we make about how they will act based on the impression we have formed [8,9].

An influential element of this process is stereotyping—the use of social identity information to develop expectations and make inferences about another's traits, intentions and actions [10]. Stereotyping is a largely automatic process, whereby social identity information is used to reduce informational load and decision time in social interaction [11,12]. This efficiency comes at a cost, however, and there is ample evidence that stereotyping introduces biases that adversely affect the fairness of social judgements and interactions [7,10,13].

The costs of this bias can be high. In job interviews, stereotype-induced bias may be particularly consequential for a person's financial and social well-being. Despite legal proscriptions against hiring on the basis of social identity, the automaticity with which we stereotype one another means that candidates with stigmatized or low-status identities (e.g. based on gender, ethnicity, pregnancy, weight or sexual preference) are frequently subjected to prejudiced decision-making resulting from the (often unintentional) application of stereotypes [14,15]. Similarly, biases in medical decision-making can have profound effects on the care a patient receives, their long-term health, and the financial burdens for them and their family members associated with protracted illness [16].

Given the social and economic costs, how can such identity-based biases be minimized in social interactions, such as job interviews, to ensure fair decisions? Given the automaticity with which stereotypes are elicited, an ideal solution is one in which cues to social identity are concealed from the interviewer entirely, without interrupting the transmission of other information relevant to the selection or evaluation process.

A real-life example of such an intervention was the introduction of 'blind auditions' to decrease gender bias in hiring by the New York Philharmonic orchestra, a historically male-dominated orchestra. To address its longstanding gender disparity, the Philharmonic's audition protocol was changed such that it required auditioners to perform behind a curtain, with women forgoing high-heels for flats so that their footsteps were indistinguishable from the men's. With visual and auditory cues related to the gender removed in this way, the employment rate for women has risen from 5% in the 1960s to 45% in the late 1990s, on course to reach gender parity [17]. Given the success of gender-blind auditions in the Philharmonic, how might similar effects be achieved for the kind of verbal communication common in other job interview settings?

Social robotics holds the promise of meeting this ideal in real-time social interactions. Telecommunication robots offer the possibility of transmitting the relevant behaviours of a real human (e.g. their words, body language and other communication cues) to another person via a robot. Unlike video-communication, telecommunication robots enable a form of telepresence that can rely on the fluency of direct dialogue *and*, for the person conversing with the robot, the sensory information of a human or humanoid shape in three-dimensional physical space. While some effects of telepresence via robots have been studied in HRI (Human Robot Interaction) research, these studies focused either on the technical aspects of telecommunication [18] or on the experience of the person operating the robot (i.e. the sense of agency and bi-location [19–21]). Yet little is known about how telepresence via robot might reduce social biases in decision-making or the social cognitive mechanisms through which such an intervention might operate. Such research is important for informing the design of systems that could serve this purpose. A guiding question for this research is: if social biases affect our decisions in assessment communication (e.g. job interviews), could candidates use suitable robotic proxies to increase social fairness? This is the basic idea behind FPC—a format for human–robot interaction that permits the transfer of communicative content in the absence of biasing social cues that should be irrelevant to the communication.

3. Fair proxy communication

FPC is a new application of social robotics technology that uses telepresence to promote equity in social interactions often fraught with bias. Given the subtlety and complexity of human social interactions, alongside the lack of precise terminology in HRI research, the communicational format of FPC has been defined previously with special attention to detail and precision [6]. For our purposes, it will suffice to mention that FPC consists of (i) a *type of communicational scenario* that is relevant to a certain type of decision, (ii) a *set-up* (illustrated in figure 1) involving the 'telepresence' of one communication partner who remotely operates a robot, (iii) the *practical ethical goal* of increasing the perceived fairness of the decision to be taken, and (iv) a certain physical and kinematic design of the robot that can fulfil this practical goal in the given scenario type for the given type of decision. Simply put, the format of FPC is employed when, in an assessment communication, the candidate to be assessed is telepresent, operating a humanoid robot that is designed as a 'neutral' human being without any social cues, in order to reduce social bias in the person conversing with the robot embodying the candidate. The candidate is directly present only via her or his utterances (although some social aspects of the acoustic information, such as pitch, intonation, etc. may be additionally masked).

FPC is novel in the sense that although robots have been used for 'telepresence' in the kind of communicational set-up described here for almost two decades [22,23], they have not yet been used with the communicational goal of reducing social bias. Nor have any robots been designed specifically with this function in mind, to our knowledge.

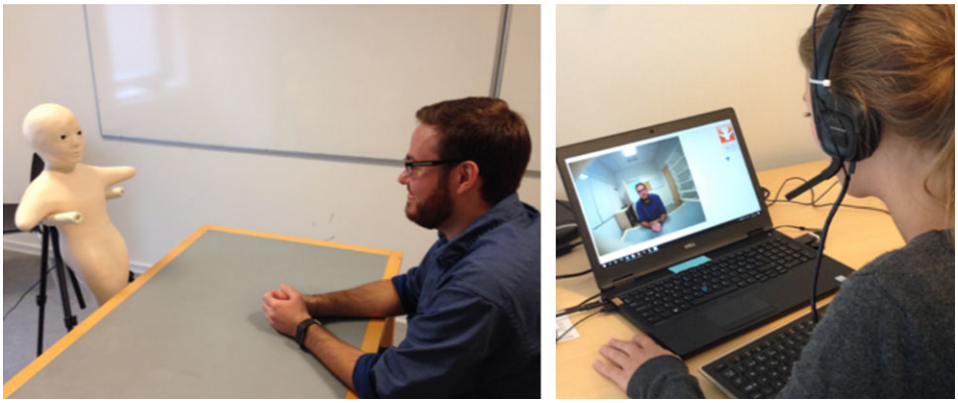

Figure 1. A job interview using FPC. The male interviewer, H_2, in the picture on the left communicates with H_1, the female job candidate in the picture on the right, via the Telenoid robot.

An example of an existing robot that might be applied in FPC is the Telenoid^{TM} R1 robot (figure 1), created by the Japanese roboticist Hiroshi Ishiguro. The Telenoid is designed to represent a 'minimal human being' without any physical indicators of social or personal identity. Ishiguro himself claims that the Telenoid is 'like an empty screen onto which specific features of the remote conversation partner can be projected'. However, such a projection is in fact unlikely to occur, according to qualitative studies undertaken by us and other research teams in Denmark [6,24,25]. Rather, the lack of perceptual identity cues appears to create a genuinely novel cognitive situation, in which the interlocutor immediately recognizes that they are interacting with *another person*, but does not automatically identify that person with any social identity category [6].

FPC can therefore be used in all situations where direct communication of one party should ideally be combined with anonymity or de-identification of the other party. FPC is particularly desirable in communication contexts where social identity bias and stereotyping are possible, and where there is an asymmetry of decisional power, such as between job candidates and employers. Furthermore, the FPC approach acknowledges that the effects of social biases on judgement can be extremely difficult to detect and control [26,27], and thus it may be more effective to design situations that preclude the activation of stereotypes in the first place rather than rely on the perceiver's ability to discount them [28].

In ongoing research, FPC is currently investigated for use in job interviews and various mediation and facilitation scenarios that can be negatively affected by the so-called mediator bias; that is, the potential bias of the mediator as perceived by the conflict parties [6]. Importantly, in the context of job interviews, FPC is not presented as a replacement for the personal interview, but as a tool that is used at stage 2 of a three-stage job interview process, following stage 1, in which anonymized CVs are screened by the company's interviewer, and prior to stage 3, which includes an in-person interview of the candidate by management. Unlike stage 1, the use of FPC in stage 2 permits a direct and dynamic interaction between the evaluator and candidate while still retaining anonymity of the candidate's pertinent social identities, prior to a final interview (stage 3) in which it may be reasonably necessary to meet in person. In effect, the anonymity of a candidate's social identities, and thus the preclusion of bias, is preserved as long as possible in the selection process. This approach provides an illustration of how FPC could be practically implemented to reduce the impact of bias in a real-life evaluative context.

4. Four preliminary concerns about fair proxy communication as a tool for social cognition research

The concept of FPC, and the kind of use case that we envision, raise a number of initial concerns regarding its usage, validity, and viability [6]. Here, we wish to highlight four concerns one might raise about the very idea of FPC as a tool for studying social bias, prefaced with a remark about potential ethical concerns.

Because it involves the use of a robot to incline a person towards debiased decision-making, FPC presents a special case of 'ethical nudging'. It is a current concern of roboethicists to investigate the lessons of larger debates on the ethics of nudging (or libertarian paternalism [29,30]) for the particular case of robot-nudging in FPC. We would like to stress, however, that while other proposals for robot-nudging discuss manipulative effects of human–robot interaction with children [31], FPC is applied in professional contexts with the explicit intention to reduce bias against under-represented group members and, in the application models we currently develop, to be followed by reflective deliberation. Furthermore, beyond its direct effects on debiasing, FPC can serve as an effective instrument to raise individual and social awareness about implicit bias while at the same time increasing equity of opportunity.

We now turn to four theoretical concerns one might raise with regard to FPC. For brevity, we refer to a robot that is designed or selected for use in FPC as an 'FPC system'.

(A) Is it possible to design a humanoid robot without *any* social identity cues? Even though the Telenoid robot may seem to qualify as an FPC system, one may question whether it fully eliminates social identity cues as opposed to merely reducing them. As has been shown in other contexts, depending on the physical and kinematic design of the robot—its shape, colour, features, and manner of interaction—as well as the context of the interaction, social robots that are not designed as FPC systems can and often do convey cues to human social identities, such as to race, ethnicity, gender, age and social class ([32–34], but see [35]). These identity cues can in turn activate concepts of ingroup or outgroup

membership, along with their associated intergroup attitudes, emotions, motivations, and any stereotype-related traits and attributes [36].

These effects must be considered when designing an FPC system or determining whether an available robotic design can serve the purposes of FPC. For example, although the Telenoid robot described above is designed to portray a blank slate without existing identity cues, its white colour could plausibly cue White racial identity, as well as ingroup or outgroup membership depending on the interaction partner's (i.e., H2 in figure 1) race [7]. Moreover, its size and shape could cue concepts of babies or children, and the effect of this identity cue could further depend on the interaction partner's identities (e.g. as an adult or parent). Human perceivers typically have a strong motivation to categorize on the basis of identity [37], and such categorizations can occur rapidly and automatically, often in the absence of awareness [7,38,39].

However, these concerns do not affect the viability of FPC in principle, nor do they rule out using the Telenoid as an FPC system. One might attempt to control for any effects of a user's social identity (e.g. of White racial identity or male gender, following the examples above), by either restricting one's sample (e.g. in a research context) or statistically adjusting for user self-identity when evaluating the results of FPC from a set of interactions (e.g. when there are multiple interviewers in a hiring procedure).

More importantly, residual social identity cues are a problem for FPC only to the extent that such cues are projected unevenly across robot interaction experiences. If a particular robot conveys the same identity cue to all interaction partners—that is, confers *equal representation*—then any effect of that cue would be held constant, and the core objective of removing individual-level bias from communications across individuals would be preserved. Equal representation in FPC is thus easily obtained when one evaluator interacts with multiple communicators (e.g. when a single manager interviews and evaluates multiple job applicants via FPC). However, if multiple evaluators are involved, it is possible that each evaluator will project somewhat different identities onto the robot. Care would be needed when aggregating evaluations to ensure that equal representation is maintained.

(B) Can cultural imagery surrounding the very concept of a robot influence the social debiasing that fair proxies seem to affect? Cultural perceptions of robots are strongly influenced by their depictions in film, anime, and consumer electronics, and these raise issues beyond those of human social identity cues noted above. From clunky contraptions (e.g. 'Robot' from *Lost in Space*) and complex machines (e.g. robots in *Transformers*), to sophisticated humanoids (e.g. Data in *Star Trek: Next Generation*) and disembodied assistants (e.g. HAL in *2001: Space Odyssey*, or Amazon's *Alexa*), robots are typically portrayed as logical computers that lack essential human qualities of emotion, intuition, and the capacity for pain. We can only assume that such conceptualizations and stereotypes may influence the way a social robot and its communicative content are perceived. To address these concerns, a social robot can be designed with a unique appearance to avoid preconceptions based on prior exposure to various kinds of robots, or with an appearance that controls for any unwanted preconceptions. For example, the design of the Telenoid robot conveys a benign appearance, and thus while it may induce a more receptive stance from the interaction partner, it minimizes more problematic associations of robots as clunky or threatening. In addition, as noted above, the key issue for FPC is that all users have the same experience with the robot (i.e. *equal representation*). And thus, although preconceptions of robots may influence the interlocutor's and other perceivers' interpretation of a robot-mediated communication, it would not systematically affect their processing of the communication if this influence is similar across users. In this case, FPC would be maintained.

(C) An additional concern regarding social identity cues is whether identity associated with the robot could interact with the stereotypes linked to the task of interest [40,41]. For example, in a job interview, would any gender cues induced by the robot affect the evaluation of candidates differently depending on a job's association with gender stereotypes (e.g. for a position as a nurse as opposed to an engineer)? The context or task or job role that frames a specific interaction may interact with social identity cues, and evaluators may judge individuals not only on the basis of identity cues themselves, but also on the fit between identity cues and stereotypes associated with the role or task the individual is required to complete. So, for gender, the effect of bias is not simply that men are preferred over women, but that individuals of a specific gender (e.g. women) will be preferred for specific roles (e.g. nursing) associated with gender stereotypes. This effect of gender-type matching in decision-making and individual-role matching is already well documented [14,42]. FPC is not equipped to remove bias linked to tasks or roles. It only functions to remove bias associated with the communicator. Assuming equal representation of any robot social identity cues, no additional bias would emerge beyond stereotypes linked to the task itself. To more fully remove bias in such cases, one would need to remove social stereotypes associated with the task or role in society.

(D) What is the role that auditory information plays in FPC? In robot-mediated FPC, linguistic information from the communicator is conveyed directly and fully, while elements of body language (facial expressions, head movements) are transmitted in a reduced fashion. Should auditory information be manipulated to be gender neutral, or standardized as male- or female-sounding? Is it possible to create a gender-neutral voice? Furthermore, other linguistic cues, such as language use and prosody, could still be transmitted, and these may also convey identity information. Moving forward, an important question will be whether the voice and other linguistic properties should be manipulated (e.g. with software) to fully standardize the communication in relation to relevant social identities.

An important factor to consider when addressing this question is the role of voice pragmatics in communicating information that might be relevant in an assessment situation. Auditory features such as word rate, pause length, prosody, changes in prosody and stuttering can convey pragmatic information about the communicator's mood state, their level of commitment to a statement or their knowledge certainty, all of which could be relevant in assessment and negotiation situations where FPC might be useful. Ideally, then, if some sort of voice modification is used in an FPC system, all relevant pragmatic information should still be transmitted in context, so as to give assessors maximal information to guide their decisions. Research on the use of FPC in these situations should therefore contribute to our knowledge of the ways in which these pragmatics operate and interact

with social identity information in relevant contexts, and this knowledge should in turn inform the design of FPC systems.

With these four preliminary concerns in view, we now turn to considering the potential for the FPC approach to advance basic research in social cognition and intergroup bias.

5. Using fair proxy communication in the social cognition laboratory

A central goal of this article is to describe how the use of social robots as 'fair proxies' can advance knowledge about mechanisms of social cognition and bias, while informing novel interventions to reduce bias. In this section, we briefly describe how this approach may be used to eliminate identity-based bias and then discuss how may be used to illuminate the sociocognitive and neural processes involved in person perception and intergroup bias.

There are in general two ways in which telecommunication robots can be used to study the elimination or mitigation of identity bias. The first major approach, which is the strategy of FPC, is to *obscure* the social identity of the communicator (H_1 in figure 1) entirely, replacing it with a design having no specific identity cues. Robots such as the Telenoid were designed to enable this approach, in comparison with robots with more pronounced identity cues. This approach has the advantage of reducing the role of social identity in the interaction as much as possible. Yet, as discussed above, it may be challenging to avoid any spontaneous associations with human groups or popular depictions of robots, and it remains unknown whether humans are able to interact with any agent without making assumptions about identity.

The second major approach is to *replace* features of the communicator's real social identity with other cues, for example, using the robot to present a female interlocutor under a male identity, or a robot identity instead of a human identity. This has the advantage of decreasing novelty while standardizing the experience across different communicators, for example, if all job applicants interact via the same robot. By providing a clear identity to the communicator, one could essentially control and equate identity classifications (and any associated assumptions or stereotypes) and satisfy the basic motivation to categorize an agent before engaging with it. We hasten to add, however, that the identity-replacement approach is an option for research purposes only; this approach goes against the letter and spirit of FPC, because it raises ethical concerns about treating specific social identities as default or more natural than others (e.g. if all candidates were presented as male) and about the erasure of marginalized identities in preference for more dominant ones. Nonetheless, these two approaches—to eradicate or to replace identity—represent two ways of using telecommunication robots for research on social cognition and bias.

In the laboratory, FPC systems can be conceived as providing a blank slate upon which to layer and manipulate various cues to social identity for the purpose of investigating basic social cognition processes. With the ability to incrementally control the display of identity information, FPC may be used to design experiments to gauge the degree to which specific identity cues contribute to biased perceptions, impressions and communicative understanding. For example, by manipulating visual features, tone of voice, prosody and body language, researchers may be able to determine the unique impact of each type of cue, as well as interactive effects between multiple cues, both on the construction of social identity and on stereotype activation and judgement bias. In this way, FPC may be used as a research tool to illuminate basic mechanisms of social categorization and the processes through which they influence judgements and actions. We explore this issue formally in relation to identity information, social identity stereotyping, and social decision-making, in §§6 and 7.

Furthermore, because FPC systems involve the representation of real human social interaction in the form of a controlled robot interface, this approach affords new opportunities to examine this type of social interaction in neuroimaging environments such as functional magnetic resonance imaging and electroencephalography. The use of FPC in neuroimaging work inspires an expanded set of theoretical questions to be examined, beginning with questions regarding the nature of human–robot interaction [1,43]. For example, to what extent is the face of a social robot, such as the Telenoid, processed like a human face, as indicated for instance by the N170 event-related potential [44] and activity in the fusiform gyrus [45]? Initial findings reveal that both human faces and humanoid robot faces (which contain human features but are clearly robots) elicit an N170 ERP response, indicating early configural face encoding [46]. However, in this research, the N170 response to robot faces is delayed, suggesting that the initial face encoding process is nevertheless reduced relative to human faces, similar to the effect observed for the visual processing of ape faces relative to human faces [47] and human outgroup faces [48,49]. Hence, an assessment of face encoding may illuminate the extent to which a social robot is perceived as human, and thus affording social identity ascriptions, particularly in the context of FPC.

A related question concerns the extent to which activations in the putative social cognition network—including the medial prefrontal cortex (mPFC [50]) and right temporo-parietal junction ([51])—and in action-observation networks [52,53] are implicated in the observation of robot actions. These networks have been previously implicated in the anthropomorphization of non-human objects (e.g. [54,55]). In the domain of human–robot interaction, research examining these questions has shown that dynamic facial expressions in humanoid robots, which have basic human facial features (eyes, mouth) yet clearly appear non-human, activate regions associated with face processing (fusiform gyrus) and the detection of agentic movement (superior temporal sulcus), to similar levels observed for dynamic human facial expressions. By comparison, regions associated with mental state inference (mPFC) and representations of social knowledge (anterior temporal lobe) were less responsive towards robot than human faces [56]. This pattern suggests that similar neural processes are involved in the perception of faces and their configural movements, but that the associated inferential processes typically engaged for humans are muted during interactions with robots. In line with these findings, other research suggests that the engagement of neural social cognition network activity depends on the degree to which a robot resembles a human [57], although there may be a point at which a close-yet-imperfect resemblance may induce social discomfort (the so-called uncanny valley effect [58]). Still other research has reported reduced activation in regions associated with affect during interactions with

humanoid robots compared with interactions with humans, suggesting diminished emotional responding and the possibility that human–robot interaction may be guided by cognitive processes more strongly than emotional processes [59]. Given prior evidence that some forms of implicit bias are amplified by intergroup anxiety [60,61], an attenuation of affect during human–robot interactions may further reduce the influence of bias.

Research has also begun to elucidate the effects of identity cues on mental state inference in human–robot interaction. Indeed, gender and race cues in the designs of virtual experiences are known to elicit social identity bias (e.g. [62–64]). Hence, it is possible that adding social identity cues to non-humanoid robots would enhance the engagement of sociocognitive inference, which in turn could affect judgements and decisions regarding robot interaction partners. Yet, less is known about the impact of robot-mediated communication like FPC on mental state inference. Existing research on mental state inferences in human software interaction has, to date, yielded mixed results. Some studies suggest that mental state inference is enhanced when the software-mediated interlocutor is believed to be human (e.g. [65]), whereas others have demonstrated similar communicative behaviours towards interlocutors believed to be a real person or a software agent (e.g. [66]).

This emerging literature reveals the growing need for a theoretical understanding of how identity-less robots, such as the Telenoid, are perceived and processed by human interaction partners in contexts like FPC. More broadly, these findings highlight the need for systematic investigation of how identity cues modulate a human's psychological experience with a robot, as well as the role of identity-based bias in this experience—issues for which FPC systems will be uniquely informative. In the following two sections, we focus on one mechanism of social cognition and discuss, in detail, how the effect of robot–human interaction on this mechanism may be investigated and implemented in the context of FPC.

6. Applying fair proxy communication to a source of implicit bias: the case of attribute substitution

One source of bias in judgement and decision-making is the use of heuristics [13,67]—information processing shortcuts that enable inferences when a perceiver has inadequate knowledge to make judgements [11,37,68]. One such heuristic is *attribute substitution* [69,70], whereby knowledge associated with a task-irrelevant but highly accessible attribute of an object is substituted for psychological states associated with a task-relevant but less accessible attribute. Attribute substitution occurs when people are faced with a complex judgement using incomplete or partially inaccessible information. A canonical example is the so-called *beautiful-is-familiar effect* [71], whereby beautiful faces are judged to be more familiar because the positive and highly accessible feeling caused by the 'beauty' attribute is substituted for the equally positive but less accessible sense of 'familiarity'.

Attribute substitution is likely to occur when three conditions are met [69], and all three of these conditions are common in the kind of scenarios in which the communicational goal of FPC is worth pursuing. The first is that the relevant target attribute (e.g. familiarity) must be transmitted unreliably or be less accessible. Relevant target attributes for evaluations and job interviews—including, for example, the ability to cope with stress, interpersonal communication skills and the ability to effectively manage expectations—are all under-defined, difficult to measure objectively and context-dependent. The second condition is that the associated but task-irrelevant attribute (e.g. beauty) must be clearly transmitted or be more easily accessible. Most forms of social identity information, especially for perceptually accessible identities such as gender and ethnicity, are highly salient. The third condition is that the substitution must not be detected by the decision maker. Although irrelevant, social identity information is strongly associated with target information via strong cultural stereotypes, which in turn may operate in the absence of awareness [11]. The result is that stereotypical attributes associated with social identity are more easily substituted for the more immediately relevant but less accessible target attributes, without the interviewer being aware of the substitution.

What is the most relevant cognitive process for explaining how and why attribute substitution readily occurs in these situations, such that there is an increased risk of implicit bias? Such attribute substitution effects may be most directly explained by the operation of semantic associative memory [72]. This system stores conceptual knowledge and the relationships between concepts in structural networks stored in the brain. At the level of social processing, semantic memory is also responsible for storing the conceptual associations that underlie implicit stereotypes [73,74]. It is this system that operates improperly when relevant target attributes are inappropriately replaced by irrelevant attributes that are more salient in the situation. In terms of its operation on cognitive function, FPC can therefore be interpreted as an intervention on semantic associative memory processes in social judgement and action.

Next, we describe a theoretical framing that may be used to develop a deeper understanding of the cognitive mechanisms through which FPC may function. We present this framework as a useful tool for developing empirical models of this process in future research.

7. Formal modelling of fair proxy communication in bias reduction research and interventions

FPC seeks to reduce biasing effects of stereotypes by limiting social identity information that might lead to attribute substitution within associative memory. Here, we present an initial attempt to theorize more fully about how such an intervention might work, to demonstrate the potential impact of such an intervention and to offer an example of how FPC might be used to test a basic cognitive explanation of implicit stereotype effects described above.

We do this in the context of a theoretical simulation of the biasing effects of the so-called think-manager/think-male implicit conception of leadership and management [75,76]. This simulation is designed to address two interrelated questions. The first is practical and concerns the importance of social identity-limiting interventions, such as FPC: what is the impact of social identity information on producing bias via attribute substitution? In other words, exactly how likely is it that attribute substitution (and therefore identity

bias) will occur under a given set of circumstances? Such specification is important, because the greater the risk of bias, the more pressing the need for the development of effective interventions of the kind discussed here.

The second question is theoretical, but relates to the first: what are the specific semantic memory mechanisms that lead to attribute substitution in such scenarios? An answer to this question is important because with more detailed knowledge of the processes involved, it will be possible to design better interventions against bias and to develop more specific hypotheses when running experiments to test those interventions. By applying a widely accepted theoretical model of associative memory for this purpose, we show in this section that even weakly held adherence to implicit stereotypes can powerfully influence our judgements of others. Such a demonstration provides a more detailed theoretical motivation for the use of FPC in reducing bias, while suggesting a framework for the use of FPC in future empirical research.

Imagine that a search committee has been asked to fill a position in management. In addition to the main job description and CV requirements, the search criteria include a set of desirable personal attributes, such as communication skills and the ability to make decisions under pressure. Consider also that committee members hold in their minds an implicit model of good leadership. It is likely that this model will associate leadership with stereotypically male attributes, including charisma and strength [14,77,78]. Assume that, as in any standard job interview situation, these attributes are easily recognized, with masculinity and strength being particularly accessible [79]. Assume also that the job criteria (e.g. decision-making skills), by contrast, are less recognizable and thus less accessible. From our discussion of attribute substitution in §6, we should expect that there is some likelihood that the committee inadvertently substitutes the salient information about the stereotypically 'male' attributes of candidates for the more relevant but less obvious target attributes listed in the job criteria, resulting in a bias towards male candidates.

A simple, productive and widely applied theoretical framework that is well suited to understanding the operation of associative memory in this task is the activation calculus from the ACT-R cognitive architecture [80]. This calculus treats semantic associative memory as a network of associations relating mental categories cued by the environment (e.g. the apparent strength of a job candidate) to categories elicited as responses (e.g. that the candidate will be a good leader). Associations between input and output categories are defined as a set of strengths S. Each output has a baseline strength B, which is independent of inputs and associations, and which is dependent on the broader context of the judgement. Within this architecture, the 'accessibility' of an attribute may be formally defined as the salience or input weight W, which denotes the influence a given input has on a given response, independent of associations in the network. With these terms in place, activation of an output category is determined by the activation equation:

$$A = B + \sum WS.$$

Conceptually stated, this equation means the following: The strength of category activation in a context is dependent on the baseline strength of that concept, plus the salience of all of the inputs presented in the environment that are associated with that concept, modulated by the strength of each of the associations in the network. The key point is that we can understand attribute substitution, and thus implicit stereotypes, in terms of the operation of this associative network.

To demonstrate, we can apply the activation equation to analyse the communicative set-up in the job interview as follows. Let us start by defining the candidate who is hired as the one who is most strongly associated with the leader/manager output concept at the end of the interviews. For the sake of simplicity, let us set this concept as the only output and fix its baseline strength at 1 (we could as easily have included other relevant job categories, such as secretary, sales assistant, accountant, etc.). The theory assumes that input categories are provided as stimuli from the environment. Thus, let us conceptualize each applicant as a cluster of attributes, with saliency values for each attribute varying between candidates. Let us assume that attributes more closely linked to appearance, such as strength, will naturally have a higher ceiling on salience than latent attributes, such as problem-solving ability.

Now let us imagine for simplicity that only two candidates are interviewed. One is a maximally strong-looking male who presents good problem-solving skills but medium interpersonal communication. Let us set the input vector for this first candidate at strength = 2, masculinity = 2, problem solving = 1 and communication = 0.5. The other candidate is a female with medium apparent strength, who presents good problem-solving skills and good communication skills. Let us set this candidates input vector as strength = 1, masculinity = 0, problem solving = 1 and communication = 1.

All that remains is to set the strength of the associative connections for each committee member. Given that they have been specified as target attributes, let us assume higher associative strength for the stated job criteria. If the committee member believes strongly and explicitly in the think-manager/think-male model, then the associative links for masculinity and strength will be strong. If they believe weakly or implicitly in the model, these links will be weaker.

With the model set up in this way, we can clearly see the role of semantic associations in determining judgement about others in job interviews and other assessment situations. This allows us to more precisely frame our practical question about the importance of FPC as an intervention against identity bias as well as our scientific question about the role of semantic memory in bias. Both questions can be reformulated more precisely in terms of the model as follows: with every parameter except the associative strengths for the irrelevant stereotypical attributes set at reasonable values, how strongly does the committee member need to be committed to the think-manager/think-male model for him/her to make an unbiased judgement of the candidates?

Figure 2*b* shows the results of this simulation. The figure suggests that even with very weak endorsement of the think-manager/think-male stereotype—that is, even when the associative strengths for the irrelevant cues are as low as 10% of the strengths for the target cues—biased judgement is more likely to occur.

This conclusion rests on an assumption that irrelevant gender cues are twice as salient as the target attributes. Let us fully relax this assumption and set the salience of the gender and strength attributes to be equal to the latent target attributes. This is equivalent to assuming that latent attributes like problem-solving skills are as accessible to a committee as are perceptible attributes like gender. Such an assumption is unrealistic [7], but useful in setting an upper bound on our results.

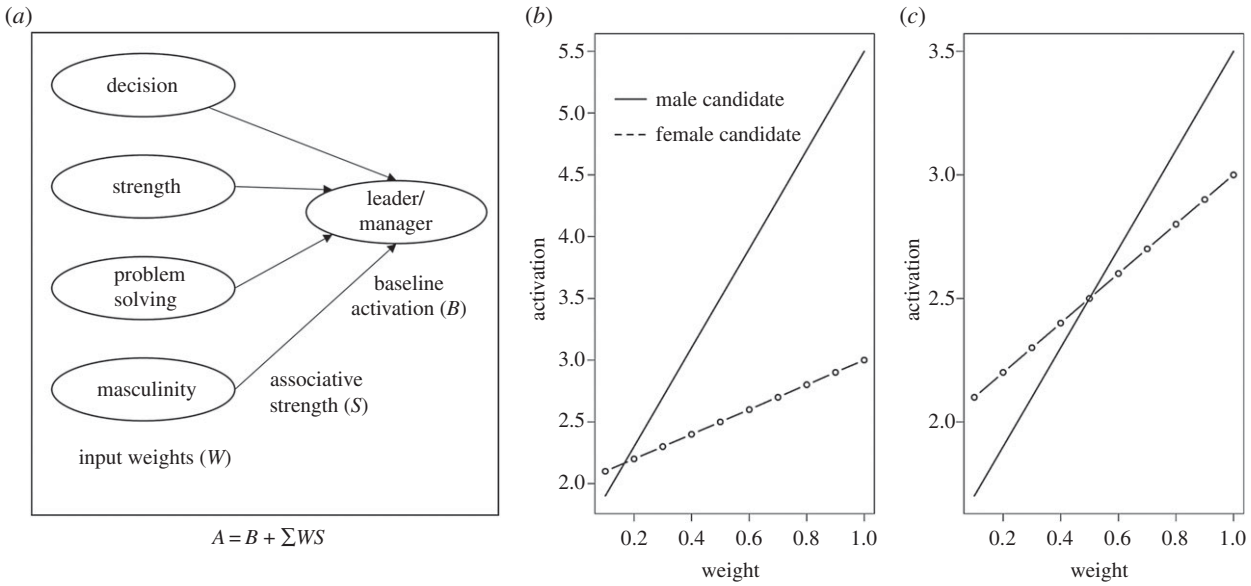

Figure 2. (*a*) Schematic illustrating the ACT-R activation calculus as a mechanism through which FPC may operate in reducing bias in a job interview context. (*b*,*c*) Simulated activation strength for the leader/manager output category, when latent target attributes for the hiring situation are less salient than irrelevant gender attributes (*b*), and when target attributes are set to the same salience level as the irrelevant attributes (*c*). The figure shows that when target attributes are less salient than the irrelevant gender cues, the committee member only begins to favour the more relevant female candidate when the associative strengths for the irrelevant attributes are set to about 10% of the associative strength of the target attributes. This suggests that even very weak endorsement of the think-manager/think-male stereotype will lead to bias. The figure also shows that even when target attributes are (unrealistically) defined as being as salient as the irrelevant gender cues, the committee member still only begins to favour the more relevant female candidate when the associative strengths for the irrelevant attributes are set to about 50% of the associative strength for the target attributes.

Figure 2*c* shows that under these conditions, the committee member begins to favour the second (female) candidate when the associative strengths for the irrelevant attributes are reduced to about 50% of the strengths for the target attributes. This suggests that even in the perfect scenario, when the salience of gender attributes is reduced to unrealistically low levels, even moderate endorsement of the think-manager/think-male bias will cause biased judgement.

This simulation offers a clear theoretical mechanism by which implicit stereotypes influence judgement, and it provides a useful justification for our expectation that FPC might be a particularly effective intervention for reducing identity bias. When the ideal conditions for attribute substitution hold—that is, when irrelevant stereotypical attributes are highly salient and target attributes are not—then only very weak associations are needed to cause bias. Such weak associations are unlikely to be explicitly endorsed, and they are also unlikely to be detected by the individual. In the absence of this knowledge, it is likely that committee members will believe they are hiring fairly, on the basis of target attributes only. By contrast, if fair proxies are used by one of the meeting participants, then salient but irrelevant attributes leading to bias may be precluded, minimizing the risk of attribute substitution. This type of formal modelling approach can be used to quantify predicted effects of FPC in such contexts, which may then be tested with behavioural experiments and compared with data from field interventions.

8. Caveats and considerations

FPC suggests a promising intervention to reduce the biasing effects of stereotypes in interactional settings as well as a promising experimental context for developing social cognitive theory. Given the mixed success of previously proposed debiasing methods (e.g. [81]), FPC has the potential to shed light on the limitations of conventional bias training and the need for alternative, complementary approaches (see also [28,82]). To be clear, our suggestion is not that existing practices, such as short-term bias training, should cease. However, the formal framework proposed in §7—which reveals that stereotype associations need only be very weakly held to have robust biasing effects—suggests that it will be difficult for conventional short-term interventions to effectively reduce bias. Our analysis suggests that a more effective strategy would be to prevent the initial activation of such associations through a procedure (i.e. FPC) that conceals irrelevant identity cues. FPC may prove particularly useful, given the increasing frequency of human–robot interactions, and thus the opportunities for using FPC to reduce bias will become increasingly ubiquitous, when compared with more idiosyncratic applications such as the blind audition procedure in orchestras. Thus, as social robots develop and their integration in the workplace expands, we expect this approach to become increasingly practical.

Research on the manipulatory effects of social robots also raises substantive ethical concerns, because such research may be used to benefit the distribution of this technology in contexts where such manipulation or 'nudging' is undesired. For this reason, value-sensitive and value-driven procedures are being developed [5,83–85] that reduce these risks by focusing research and development in social robotics from the very beginning on the protection and enhancement of ethical values. The concept of FPC is the result of such a value-driven design approach.

9. Conclusion

Recent advances in the capabilities of programmable robots represent a new frontier in the science of the human mind. Here, with our focus on FPC, we have illustrated how

applications of social robotics may be used to elucidate the cognitive mechanisms of implicit bias while also promoting the practical goal of enhancing social justice. We view FPC as an integral component of a broad praxis of bias reduction that includes individual and public reflections on the effects of FPC. While a fair proxy may decrease the probability that attribute substitution occurs in a given context, it cannot change the propensity of individuals to rely on stereotypical identity attributes when evaluating others more generally. By masking the attributes that elicit implicit bias, rather than changing the stereotypes on which biased attribute substitution is based, fair proxies provide only a partial, contextually bounded and short-term solution to the systemic problem of discrimination. High-level social structures, such as culture and institutions, must also be changed if we are to properly deal with the deeper problems of discrimination. Structures change slowly, however, and our analysis suggests that FPC may provide a promising framework for building tools to increase equity in the more immediate future.

Data accessibility. This article has no additional data.

Competing interests. We declare we have no competing interests.

Funding. Work on this article was supported by the Carlsberg Foundation Semper Ardens Grant (CF-2016004) to J.Se. and The Netherlands Organisation for Scientific Research (VICI 016.185.058) to D.M.A.

References

1. Hortensius R, Cross ES. 2018 From automata to animate beings: the scope and limits of attributing socialness to artificial agents. *Ann. NY Acad. Sci.* **1426**, 93. (doi:10.1111/nyas.13727)
2. Fong T, Nourbakhsh I, Dautenhahn K. 2003 A survey of socially interactive robots. *Robot. Auton. Syst.* **42**, 143–166. (doi:10.1016/S0921-8890(02)00372-X)
3. Breazeal C. 2003 Toward sociable robots. *Robot. Auton. Syst.* **42**, 167–175. (doi:10.1016/S0921-8890(02)00373-1)
4. Wykowska A, Chaminade T, Cheng G. 2016 Embodied artificial agents for understanding human social cognition. *Phil. Trans. R. Soc. B* **371**, 20150375. (doi:10.1098/rstb.2015.0375)
5. Seibt J. 2016 Integrative social robotics—a new method paradigm to solve the description problem and the regulation problem? In *What social robots can and should do. Proc. Robophilosophy /TRANSOR, Aarhus, October 2016* (eds J Seibt, M Nørskov, S Andersen), pp. 104–114. Amsterdam, The Netherlands: IOS Press.
6. Seibt J, Vestergaard C. 2018 Fair proxy communication: using social robots to modify the mechanisms of implicit social cognition. *Res. Ideas Outcomes* **4**, e31827. (doi:10.3897/rio.4.e31827)
7. Kawakami K, Amodio DM, Hugenberg K. 2017 Intergroup perception and cognition: an integrative framework for understanding the causes and consequences of social categorization. *Adv. Exp. Soc. Psychol.* **55**, 1–80. (doi:10.1016/bs.aesp.2016.10.001)
8. Asch SE. 1946 Forming impressions of personality. *J. Abnorm. Soc. Psychol.* **41**, 258. (doi:10.1037/h0055756)
9. Macrae CN, Bodenhausen GV. 2000 Social cognition: thinking categorically about others. *Annu. Rev. Psychol.* **51**, 93–120. (doi:10.1146/annurev.psych.51.1.93)
10. Fiske ST. 1998 Stereotyping, prejudice, and discrimination. In *The handbook of social psychology* (eds DT Gilbert, ST Fiske, G Lindzey), vol. 2, pp. 357–411. New York, NY: McGraw-Hill.
11. Devine PG. 1989 Stereotypes and prejudice: their automatic and controlled components. *J. Pers. Soc. Psychol.* **56**, 5–18. (doi:10.1037/0022-3514.56.1.5)
12. Taylor SE, Crocker J. 1981 Schematic bases of social information processing. In *Social cognition* (eds ET Higgins, CP Herman, M Zanna), pp. 89–134. Hillsdale, NJ: Lawrence Erlbaum.
13. Macrae CN, Milne AB, Bodenhausen GV. 1994 Stereotypes as energy-saving devices: a peek inside the cognitive toolbox. *J. Pers. Soc. Psychol.* **66**, 397–407. (doi:10.1037/0022-3514.66.1.37)
14. Heilman ME. 2012 Gender stereotypes and workplace bias. *Res. Org. Behav.* **32**, 113–135. (doi:10.1016/j.riob.2012.11.003)
15. Macan, T, Merritt S. 2011 Actions speak too: uncovering possible implicit and explicit discrimination in the employment interview process. *Int. Rev. Ind. Org. Psychol.* **26**, 293–337. (doi:10.1002/9781119992592.ch8)
16. Green AR, Carney DR, Pallin DJ, Ngo LH, Raymond KL, Lezzoni LI, Banaji MR. 2007 Implicit bias among physicians and its prediction of thromolysis decisions for black and white patients. *J. Gen. Intern. Med.* **22**, 1231–1238. (doi:10.1007/s11606-007-0258-5)
17. Goldin C, Rouse C. 2000 Orchestrating impartiality: the impact of 'blind' auditions on female musicians. *Am. Econ. Rev.* **90**, 715–741. (doi:10.1257/aer.90.4.715)
18. Lepage P, Létourneau D, Hamel M, Brière S, Corriveau H, Tousignant M, Michaud F. 2016 Telehomecare telecommunication framework—from remote patient monitoring to video visits and robot telepresence. In *Proc. 2016 IEEE 38th Ann. Int. Conf. Engineering in Medicine and Biology Society (EMBC)*, pp. 3269–3272. IEEE.
19. Aymerich-Franch L, Petit D, Ganesh G, Kheddar A. 2015 Embodiment of a humanoid robot is preserved during partial and delayed control. In *Proc. 2015 IEEE Int. Workshop on Advanced Robotics and its Social Impacts* (*ARSO*), pp. 1–5. IEEE.
20. Seibt J, Nørskov M. 2012 'Embodying' the internet: towards the moral self via communication robots? *Philos. Technol.* **25**, 285–307. (doi:10.1007/s13347-012-0064-9)
21. Aymerich-Franch L, Petit D, Ganesh G, Kheddar A. 2016 The second me: seeing the real body during humanoid robot embodiment produces an illusion of bi-location. *Conscious Cogn.* **46**, 99–109. (doi:10.1016/j.concog.2016.09.017)
22. Ogata T, Matsuyama Y, Komiya T, Ida M, Noda K, Sugano S. 2000 Development of emotional communication robot: WAMOEBA-2R-experimental evaluation of the emotional communication between robots and humans. In *Proc. 2000 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2000)*, vol. 1, pp. 175–180. IEEE.
23. Kato S, Ohshiro S, Itoh H, Kimura K. 2004 Development of a communication robot ifbot. Presented at the *Proc. ICRA'04, 2004 IEEE Int. Conf. on Robotics and Automation*, vol. 1, pp. 697–702. IEEE.
24. Yamazaki R, Nishio S, Ishiguro H, Nørskov M, Ishiguro N, Balistreri G. 2012 Social acceptance of a teleoperated android: field study on elderly's engagement with an embodied communication medium in Denmark. In *Int. Conf. on Social Robotics*, pp. 428–437. Berlin, Germany: Springer.
25. Damholdt MF, Nørskov M, Yamazaki R, Hakli R, Hansen CV, Vestergaard C, Seibt J. 2015 Attitudinal change in elderly citizens toward social robots: the role of personality traits and beliefs about robot functionality. *Front. Psychol.* **6**, 1701. (doi:10.3389/fpsyg.2015.01701)
26. Amodio DM. 2011 Self-regulation in intergroup relations: a social neuroscience framework. In *Social neuroscience: toward understanding the underpinnings of the social mind* (eds A Todorov, ST Fiske, D Prentice), pp. 101–122. New York, NY: Oxford University Press.
27. Amodio DM, Harmon-Jones E, Devine PG, Curtin JJ, Hartley SL, Covert AE. 2004 Neural signals for the detection of unintentional race bias. *Psychol. Sci.* **15**, 88–93. (doi:10.1111/j.0963-7214.2004.01502003.x)
28. Amodio DM, Swencionis JK. 2018 Proactive control of implicit bias: a theoretical model and implications for behavior change. *J. Pers. Soc. Psychol.* **115**, 255–275. (doi:10.1037/pspi0000128)
29. Sunstein CR, Thaler RH. 2003 Libertarian paternalism is not an oxymoron. *Univ. Chicago Law Rev.* **70**, 1159–1202. (doi:10.2307/1600573)

30. Thaler RH, Sunstein CR. 2009 *Nudge: improving decisions about health, wealth and happiness*, 2nd edn. New York, NY: Penguin Books.
31. Borenstein J, Arkin R. 2016 Robotic nudges: the ethics of engineering a more socially just human being. *Sci. Eng. Ethics* **22**, 31–46. (doi:10.1007/s11948-015-9636-2)
32. Bartneck C, Yogeeswaran K, Ser QM, Woodward G, Sparrow R, Wang S, Eyssel F. 2018 Robots and racism. In *Proc. 2018 ACM/IEEE Int. Conf. on Human–Robot Interaction, Chicago, 5–8 March 2018* (eds T Kanda, G Hoffman, S Šabanović, A Tapus), pp. 196–204. New York, NY: ACM. (doi:10.1145/3171221.3171260)
33. Strait MK, Floerke VA, Ju W, Maddox K, Remedios JD, Jung MF, Urry HL. 2017 Understanding the uncanny: both atypical features and category ambiguity provoke aversion toward humanlike robots. *Front. Psychol.* **8**, 1366. (doi:10.3389/fpsyg.2017.01366)
34. Howard A, Borenstein J. 2018 The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Sci. Eng. Ethics* **24**, 1521–1536. (doi:10.1007/s11948-017-9975-2)
35. Ogunyale T, De'Aira B, Ayanna H. 2018 Does removing stereotype priming remove bias? A pilot human-robot interaction study. *arXiv* (http://arxiv.ord/abs/1807.00948v1).
36. Eyssel F, Kuchenbrandt D. 2012 Social categorization of social robots: anthropomorphism as a function of robot group membership. *Br. J. Soc. Psychol.* **51**, 724–731. (doi:10.1111/j.2044-8309.2011.02082.x)
37. Allport GW. 1954 *The nature of prejudice*. New York, NY: Addison-Wesley.
38. Freeman JB, Johnson KL. 2016 More than meets the eye: split-second social perception. *Trends Cogn. Sci.* **20**, 362–374. (doi:10.1016/j.tics.2016.03.003)
39. Gilbert DT, Hixon JG. 1991 The trouble of thinking: activation and application of stereotypic beliefs. *J. Pers. Soc. Psychol.* **60**, 509–517. (doi:10.1037/0022-3514.60.4.509)
40. Eyssel F, Hegel F. 2012 (S)he's got the look: gender stereotyping of robots 1. *J. Appl. Soc. Psychol.* **42**, 2213–2230. (doi:10.1111/j.1559-1816.2012.00937.x)
41. Kuchenbrandt D, Häring M, Eichberg J, Eyssel F, André, E. 2014 Keep an eye on the task! How gender typicality of tasks influence human–robot interactions. *Int. J. Soc. Robot.* **6**, 417–427. (doi:10.1007/s12369-014-0244-0)
42. Bertrand M, Mullainathan S. 2004 Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* **94**, 991–1013. (doi:10.1257/0002828042002561)
43. Wiese E, Metta G, Wykowska A. 2017 Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Front. Psychol.* **8**, 1663. (doi:10.3389/fpsyg.2017.01663)
44. Bentin S, Allison T, Puce A, Perez E, McCarthy G. 1996 Electrophysiological studies of face perception in humans. *J. Cogn. Neurosci.* **8**, 551–565. (doi:10.1162/jocn.1996.8.6.551)
45. Kanwisher N, McDermott J, Chun MM. 1997 The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311. (doi:10.1523/JNEUROSCI.17-11-04302.1997)
46. Dubal S, Foucher A, Jouvent R, Nadel J. 2010 Human brain spots emotion in non-humanoid robots. *Soc. Cogn. Affect. Neurosci.* **6**, 90–97. (doi:10.1093/scan/nsq019)
47. Carmel D, Bentin S. 2002 Domain specificity versus expertise: factors influencing distinct processing of faces. *Cognition* **83**, 1–29. (doi:10.1016/S0010-0277(01)00162-7)
48. Ratner KG, Amodio DM. 2013 Seeing 'us vs. them': minimal group effects on the neural encoding of faces. *J. Exp. Soc. Psychol.* **49**, 298–301. (doi:10.1016/j.jesp.2012.10.017)
49. Schmid PC, Amodio DM. 2017 Power effects on implicit prejudice and stereotyping: the role of intergroup face processing. *Soc. Neurosci.* **12**, 218–231. (doi:10.1080/17470919.2016.1144647)
50. Amodio DM, Frith CD. 2006 Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* **7**, 268–277. (doi:10.1038/nrn1884)
51. Saxe R, Kanwisher N. 2003 People thinking about thinking people: the role of the temporo-parietal junction in 'theory of mind'. *Neuroimage* **19**, 1835–1842. (doi:10.1016/S1053-8119(03)00230-1)
52. Cross ES, Liepelt R, Hamilton AF, Parkinson J, Ramsey R, Stadler W, Prinz W. 2012 Robotic movement preferentially engages the action observation network. *Hum. Brain Mapp.* **33**, 2238–2254. (doi:10.1002/hbm.21361)
53. Gallese V, Keysers C, Rizzolatti G. 2004 A unifying view of the basis of social cognition. *Trends Cogn. Sci.* **8**, 396–403. (doi:10.1016/j.tics.2004.07.002)
54. Cullen H, Kanai R, Bahrami B, Rees G. 2013 Individual differences in anthropomorphic attributions and human brain structure. *Soc. Cogn. Affect. Neurosci.* **9**, 1276–1280. (doi:10.1093/scan/nst109)
55. Waytz A, Cacioppo J, Epley N. 2010 Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* **5**, 219–232. (doi:10.1177/1745691610369336)
56. Gobbini MI *et al.* 2011 Distinct neural systems involved in agency and animacy detection. *J. Cogn. Neurosci.* **23**, 1911–1920. (doi:10.1162/jocn.2010.21574)
57. Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T. 2008 Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS ONE* **3**, e2597. (doi:10.1371/journal.pone.0002597)
58. MacDorman KF, Green RD, Ho CC, Koch CT. 2009 Too real for comfort? Uncanny responses to computer generated faces. *Comput. Hum. Behav.* **25**, 695–710. (doi:10.1016/j.chb.2008.12.026)
59. Chaminade T *et al.* 2010 Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. *PLoS ONE* **5**, e11577. (doi:10.1371/journal.pone.0011577)
60. Amodio DM, Hamilton HK. 2012 Intergroup anxiety effects on implicit racial evaluation and stereotyping. *Emotion* **12**, 1273–1280. (doi:10.1037/a0029016)
61. Ofan RH, Rubin N, Amodio DM. 2014 Situation-based social anxiety enhances the neural encoding of faces: evidence from an intergroup context. *Soc. Cogn. Affect. Neurosci.* **9**, 1055–1061. (doi:10.1093/scan/nst087)
62. Nass, C, Moon, Y, Green N. 1997 Are machines gender neutral? Gender stereotypic responses to computers with voices. *J. Appl. Soc. Psychol.* **27**, 864–876. (doi:10.1111/j.1559-1816.1997.tb00275.x)
63. Lee EJ, Nass C, Brave S. 2000 Can computer-generated speech have gender? An experimental test of gender stereotype. In *CHI'00 Extended Abstracts on Human Factors in Computing Systems*, pp. 289–290. IEEE.
64. Groom V, Bailenson JN, Nass C. 2009 The influence of racial embodiment on racial bias in immersive virtual environments. *Soc. Influence* **4**, 231–248. (doi:10.1080/15534510802643750)
65. Gallagher HL, Jack AI, Roepstorff, A, Frith CD. 2002 Imaging the intentional stance in a competitive game. *Neuroimage* **16**, 814–821. (doi:10.1006/nimg.2002.1117)
66. Branigan HP, Pickering MJ, Pearson, J, McLean JF, Nass C. 2003 Syntactic alignment between computers and people: the role of belief about mental states. In *Proc. 25th Ann. Conf. Cognitive Science Society*, Boston, *MA, 31 July–2 August 2003* (eds R Alterman, D Kirsh), pp. 186–191. Mahwah, NJ: Lawrence Erlbaum Associates.
67. Tversky A, Kahneman D. 1974 Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131. (doi:10.1126/science.185.4157.1124)
68. Fiske ST, Neuberg SL. 1990 A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology*, vol. 23, pp. 1–74. New York, NY: Academic Press.
69. Kahneman D, Frederick S. 2002 Representativeness revisited: attribute substitution in intuitive judgment. In *Heuristics and biases: the psychology of intuitive judgment* (eds T Gilovich, DW Griffin, D Kahneman), pp. 49–81. Cambridge, UK: Cambridge University Press.
70. Kahneman D, Shane F. 2004 Attribute substitution in intuitive judgment. In *Models of man: essays in memory of Herbert A. Simon* (eds M Augier, JG March), pp. 411–432. Cambridge, MA: MIT Press.
71. Monin B. 2003 The warm glow heuristic: when liking leads to familiarity. *J. Pers. Soc. Psychol.* **85**, 1035–1048. (doi:10.1037/0022-3514.85.6.1035)
72. Anderson JR, Bower GH. 1973 *Human associative memory*. Washington, DC: Winston and Sons.
73. Amodio DM. 2019 Social Cognition 2.0: an interactive memory systems account. *Trends Cogn. Sci.* **23**, 21–33. (doi:10.1016/j.tics.2018.10.002)
74. Amodio DM, Ratner KG. 2011 A memory systems model of implicit social cognition. *Curr. Dir. Psychol. Sci.* **20**, 143–148. (doi:10.1177/0963721411408562)

75. Heilman ME, Block CJ, Martell RF, Simon MC. 1989 Has anything changed? Current characterizations of men, women, and managers. *J. Appl. Psychol.* **74**, 935–942. (doi:10.1037/0021-9010.74.6.935)
76. Schein VE, Davidson MJ. 1993 Think manager, think male. *Manage. Dev. Rev.* **6**, 24–29. (doi:10.1108/EUM0000000000738)
77. Ceci SJ, Williams WM. 2015 Women have substantial advantage in STEM faculty hiring, except when competing against more-accomplished men. *Front. Psychol.* **6**, 1532. (doi:10.3389/fpsyg.2015.01532)
78. Offerman LR, Coates MR. 2018 Implicit theories of leadership: stability and change over two decades. *Leadersh. Q.* **29**, 513–522. (doi:10.1016/j.leaqua.2017.12.003)
79. Rudman LA, Greenwald AG, McGhee DE. 2001 Implicit self-concept and evaluative implicit gender stereotypes: self and ingroup share desirable traits. *Pers. Soc. Psychol. Bull.* **27**, 1164–1178. (doi:10.1177/0146167201279009)
80. Anderson JR. 2009 *How can the human mind occur in the physical universe?* Oxford, UK: Oxford University Press.
81. Lai CK *et al.* 2014 Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *J. Exp. Psychol. Gen.* **143**, 1765–1785. (doi:10.1037/a0036260)
82. Amodio DM, Devine PG. 2005 Changing prejudice: the effects of persuasion on implicit and explicit forms of race bias. In *Persuasion: psychological insights and perspectives* (eds TC Brock, MC Green), 2nd edn, pp. 249–280. Thousand Oaks, CA: Sage Publications.
83. Van Wynsberghe A. 2012 Designing robots for care: care centered value-sensitive design. *Sci. Eng. Ethics* **19**, 407–433. (doi:10.1007/s11948-011-9343-6)
84. Van den Hoven J, Vermaas P, Ivan de Poel I. 2015 Design for values—Introduction. In *Handbook for ethics, values, and technological design* (eds PE Vermaas, I van de Poel), pp. 1–9. New York, NY: Springer.
85. Seibt, J, Damholdt MF, Vestergaard C. 2018 Five principles of integrative social robotics. In *Envisioning robots in society—power, politics, and public space* (eds M Coeckelbergh, J Loh, M Funk, J Seibt, M Nørskov), pp. 28–43. Amsterdam, The Netherlands: IOS Press.