Research Ideas and Outcomes 4: e31827
doi: 10.3897/rio.4.e31827

Research Idea

Fair Proxy Communication: Using Social Robots to Modify the Mechanisms of Implicit Social Cognition

Johanna Seibt‡, Christina Vestergaard‡

‡ Aarhus University, Aarhus, Denmark

Corresponding author: Johanna Seibt (filseibt@cas.au.dk)

Reviewable v1

Received: 21 Nov 2018 | Published: 27 Nov 2018

Citation: Seibt J, Vestergaard C (2018) Fair Proxy Communication: Using Social Robots to Modify the Mechanisms of Implicit Social Cognition. Research Ideas and Outcomes 4: e31827. https://doi.org/10.3897/rio.4.e31827

Abstract

This article introduces a new communicational format called Fair Proxy Communication. Fair Proxy Communication is a specific communicational setting in which a teleoperated robot is used to remove perceptual cues of implicit biases in order to increase the perceived fairness of decision-related communications. The envisaged practical applications of Fair Proxy Communication range from assessment communication (e.g. job interviews at Affirmative Action Employers) to conflict mediation, negotiation and other communication scenarios that require direct dialogue but where decision-making maybe negatively affected by implicit social biases. The theoretical significance of Fair Proxy Communication pertains primarily to the investigation of 'mechanisms' of implicit social cognition in neuropsychology, but this new communicational format also raises many research questions for the fields of organisational psychology, negotiation and conflict research and business ethics. Fair Proxy Communication is currently investigated by an interdisciplinary research team at Aarhus University, Denmark.

Keywords

communication, perceptual bias, discrimination, integrative social robotics, telepresence, job interview, conflict facilitation, fairness, mediator neutrality, social cognition

Background: Integrative Social Robotics

The aim of this paper is to introduce the communicational format called "Fair Proxy Communication". The basic idea of Fair Proxy Communication was conceived by the first author in 2014, empirically explored by the second author in a pilot study at a Danish professional school in early 2015 and further developed by both authors. Fair Proxy Communication has been described in greater detail in 2016 in the context of a grant application for *Integrative Social Robotics*, a new approach to research, design and development of applications of robots in social interaction contexts. Even though the basic idea of Fair Proxy Communication is independent of this approach, it also can serve as a prime illustration of its potential benefits. In fact, the requirements for Fair Proxy Communication, as defined below, will become clearer if we introduce it from within its larger motivational context and begin with a brief sketch of background of *Integrative Social Robotics.*

"Social robotics" and "Human-Robot-Interaction Studies" are fairly young interdisciplinary research areas exploring the phenomena of human interactions with so-called 'social robots.' So far, both areas have been conducted with limited interdisciplinary scope, involving mainly robotics, psychology (developmental psychology but also autism research), geronotology and education science, but the Humanities and social sciences are not yet fully involved. This means in effect that, so far, deeply disruptive technologies have been developed and investigated without involving all relevant expertise. Given that so-called 'social' robots are to engage humans in new socio-cultural interactions, it seems irresponsible and ultimately counterproductive to disregard the expertise of the Humanities. *Integrative Social Robotics* (Seibt 2016a, Seibt 2016b, Seibt et al. 2018) is a new approach or 'paradigm' for the research, design and development process in social robotics that involves Humanities expertise from the very beginning and throughout. This new approach systematically combines social robotics research with research on socio-cultural practices and values, as undertaken in the Humanities (anthropology, ethics, value theory, education research, linguistics and communication studies, phenomenology, ontology, knowledge representation, epistemology) and the Human and Social Sciences (psychology, cognitive science, sociology and management).

Integrative Social Robotics is a targeted response to growing concerns, expressed both within the robotics community as well as in the professional and public debate on socio-cultural and ethical values, that an unregulated social robotics industry may create profound and possibly negative cultural changes (see e.g. Turkle 2011, Nourbakhsh 2013, Dumouchel and Damiano 2017). According to the current set-up for the production of social robotics applications, the research, design and development process in social robotics proceeds largely unencumbered by interactions with professional research in ethics, value-theory or empirical studies of socio-cultural practices; normative considerations enter at best *after* the technology is developed and ready-made products are to be *selected* for use by policy-makers and law-givers, who turn to ethics councils and the empirical Humanities to gauge the socio-cultural implications of the relevant applications.

This serial arrangement — first development, then professional evaluation of socio-cultural and ethical significance and finally policy and legal regulation — has two crucial drawbacks. On the one hand, due to the mentioned sequentialisation, research on the cultural-ethical implications and commercial potential of social robotics applications currently is lagging far behind the rapid developments in robot technology and the advice that policy- and law-makers can receive from national ethical councils is not always fully informed about the technology. On the other hand, as long as the methods and categories of value and social interaction research in the Humanities are not included in the interdisciplinary scope of HRI — currently mainly consisting of quantitative studies in psychology and sociology — the research, design and development process in social robotics misses out on important resources for innovation and anticipatory adjustments to expected ethical and legal regulation.

In contrast, according to the approach of *Integrative Social Robotics*, Humanities research on ethical, conceptual and socio-cultural norms and values *is both informed by and provides information on on* all stages of the research, design and development process of social robotics. Since so-called 'social' robots are no longer tools but interfere with the sphere of human social interactions at the (preconscious and conscious) semiotic level of social agency, *Integrative Social Robotics* proposes that research, design and development of social robotics applications should be joined, from the very beginning, with professional research in disciplines whose concepts and methods have been designed to explore, empirically and hermeneutically, the profound complexity of human social interactions and the cultural values constituted by these practices.

More concretely, *Integrative Social Robotics* operates with five methodological principles (Seibt et al. 2018). The first four principles:

1. call for full-scope interdisciplinary expertise as required by the envisaged social robotics application and
2. state three requirements that establish a sufficiently sophisticated or careful understanding of social interactions.

Of primary interest in the present context is the fifth principle, the "values first principle" which demands that social robotics applications should be developed with the goal of preserving or enhancing a value that has top rank in a given axiological system (ethical or sociocultural values are typically top rank values). This demand for a strictly value-driven approach is a reinforcement of cognate design principles calling for "value-sensitive" technology development (Friedman et al. 1997) or "design for values" (Van den Hoven 2005). Importantly, however, the "values first principle" also includes a "non-replacement maxim": "social robots may only do what humans should but cannot do" (ibid.). That is, the "values first principle" forces developers of social robotics applications to identify legitimate developmental targets by means of the question: 'Is there a high-ranking (moral) value that, in the given context and given certain constraints C, *cannot* be realised by human-human interaction but *can* be realised by means of a human-robot social* interaction?'. The constraints C may relate to material aspects (e.g.humans cannot be exposed to radioactive radiation or cannot run on solar energy) or to more subtle features of kinematics and

appearance (e.g. humans typically cannot repeat actions precisely and indefinitely and humans typically cannot fail to display social identity cues such as gender, ethnicity, race, age etc.).

The development of Fair Proxy Communication, the application of which we will describe in the following, is a prime illustration of the "values first principle". The non-replacement maxim demands that we identify suitable targets for social robotics applications by considering which of our top values could be supported by agents and can do what human agents cannot do. In the case of Fair Proxy Communication, the top values pursued are perceived fairness (in communication), as well as the values of social equality and/or peace as these depend on perceived fairness. Guided by the question: 'how can we enhance perceived fairness in communication – and thereby enhance social equality and/or peace–making good use of features that are unique to robots?', the first author, a philosopher, conceived of the idea of Fair Proxy Communication by combining:

1. results from experimental research in neuropsychology on implicit social cognition and the generation of perceptual biases;
2. results from qualitative studies undertaken in Denmark, using the telecommunication android robot “Telenoid R1”, developed by the Japanese robotics lab “ATR /Hiroshi Ishiguro Laboratories,” Kyoto, Japan; and
3. results of the intended application areas, especially conflict research and research on the role of perceptual biases in assessment and selection contexts.

The second author, a specialist in conflict research and anthropology, refined the initial idea and conducted first qualitative research using the Telenoid R1 robot (created by Hiroshi Ishiguro, ATR, Japan). Based on these pilot studies, we could establish Fair Proxy Communication as a viable research target for an interdisciplinary team committed to Integrative Social Robotics (26 researchers from 14 disciplines), for it appeared that certain teleoperated communication robots can "do what humans should but cannot do," namely,

1. to be physically present as three-dimensional interlocutors without displaying any cues that support the formation of perceptual biases and
2. thereby enhancing the perceived fairness of communications where perceptual biases, negative or positive, tend to lead to prejudiced decision-making.

In the following sections, we will define the idea of Fair Proxy Communication more precisely, describe its practical and theoretical objectives and explain the current implementation of several research lines exploring Fair Proxy Implementation.

Definition of Fair Proxy Communication

The field of Human-Robot Interaction Studies, a relatively new multidisciplinary research area, is still in need of more precise descriptive terminology. For this reason, we wish to offer here a fairly detailed definition of the term "Fair Proxy Communication", which is short-hand for "fairness-enhancing communication with robotic proxies" and hereafter abbreviated as 'FPC'. This is a tentative definition that we may need to adjust in the course of our empirical research, but for such adjustments, it is important to have an initial reference point with a precise definition.

Roughly speaking, FPC labels a specific communicational format that involves the ‘telepresence’ of one communication partner who remotely operates a robot with a special set of affordances and a communicational setting (such as job interviews) where perceptual biases may lead to unfair decisions. Before we present a more precise definition, let us offer an illustration; consider Figs 1, 2.

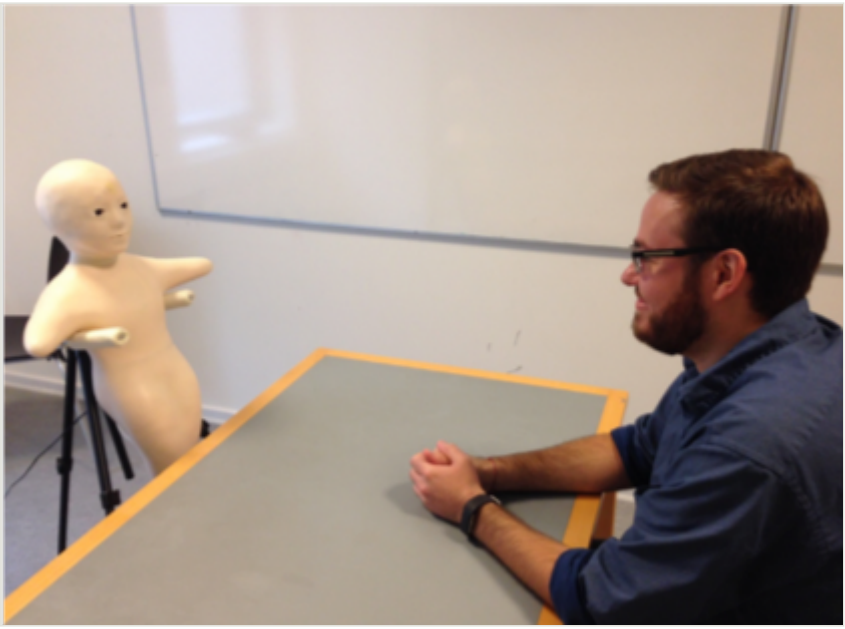

Figure 1. doi

A job interview using FPC, from the perspective of the interviewer. The male interviewer (in the definition: H_2) communicates via the Telenoid R4 robot with H_1, a female job candidate as shown in Fig. 2.

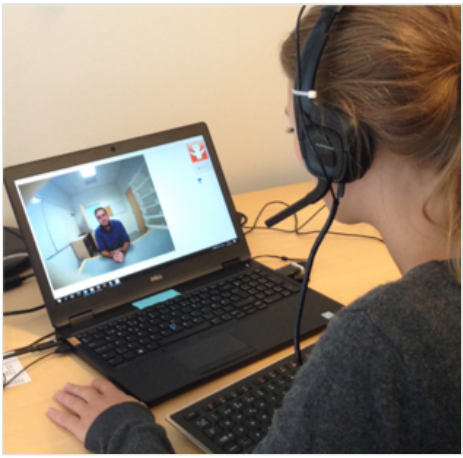

Figure 2. doi

A job interview using FPC, from the perspective of the job candidate. A female candidate (in the definition D1-1: H1) operates her robotic proxy, a Telenoid R4 robot, while she communicates with the male interviewer. Her head movements, lip movements and speech are translated directly to the robot, either via a kinetic sensor on a headset or by a facial reading programme; her voice may or may not be morphed to mask gender (see section "Practical Significance" below). The camera that projects the interviewer on to her computer screen is in the eyes of the robot –thus, in contrast to a skype session, she looks into the interviewer's eyes when facing the interviewer.

The illustrated scenario is an instantiation of the general definition of FPC, which we state as follows:

Definition (Def1): FPC refers to a specific form of communication, which consists of a *communication scenario suited for FPC* (D1-1), a condition describing the physical *set-up of FPC* (D1-2), a condition describing the *practical-ethical goal* that is to be achieved via FPC (D1-3) and means by which the goal is achieved (D1-4).

1. (Def 1-1) S is a ***communication scenario suited for FPC*** if, in S, two (or more) people, H_1 and H_2 (or H_i), are conversing with each other at close range (at a distance between about 80-200 cm), for the sake of a communicational purpose that is related to a decision D about H_1 by H_2 (or any of the H_i).
2. (Def 1-2) S is ***set up for FPC*** just in case the human interlocutor H_1, is replaced by a robotic proxy that is remotely operated by H_1. The functionalities of the remote operation are those that are sufficient and necessary to realise all communicative capacities of H_1 that are relevant for D.

3. (Def 1-3) S has been modified as described in (Def 1-1) for the sake of the ***practical-ethical goal*** of increasing the perceived fairness of S relative to D, i.e.of ensuring that H_2 (or any of the H_i) is not provided with perceptual cues (pertaining to gender, age, ethnicity, race etc.) that both H1 and H2 (and any of the H_i) know to be associated with negative or positive biases relevant for D.

4. (Def 1-4) The robotic proxy has a (physical and kinematic) ***design*** PKD that fulfils the practical-ethical goal as specified in (Def 1-2).

We supplement definition (D1) with the following points of explanatory commentary.

Comment 1: Condition (Def 1-1) requires that, for any concrete application of FPC, it needs to be clarified explicitly which decision is the target of the application. In the given illustration of the job interview, the decision mentioned in condition (Def 1-1) may be the decision to commit to a definite evaluation of whether the candidate is suited for the position or it may be the decision that is based on the assessment, namely, to hire the candidate or admit her to the next step in the selection process etc. This is crucially important for further research on FPC since it is one thing to investigate whether the removal of perceptual cues for bias achieved by FPC can alter the evaluations of job candidates and another to investigate whether it also can change actual hiring rates. Besides job interviews, there are many other suitable communication scenarios involving short-term or long-term decision-making that can be negatively affected by perceptual biases, such as oral examinations, personal meetings for application for credit, rental property or other asymmetric business transactions. Note also that the fact that H_2's decision about H_1 may merely be the decision that H_1 is trustworthy and provides good counsel – in the following section, we describe an application of FPC where conflict facilitators use robotic proxies in order to circumvent so-called "perceived mediator bias".

Comment 2: According to (Def1-1) and (Def1-2), the decisional power in S should be asymmetric and only the person that is decided upon, H_1, is telepresent in S via robotic proxy. Thus, asymmetric decisional power in S is tied to asymmetric telepresence. One might argue, however, that perceived fairness of S might be even further increased if the decision-maker H_2 also is telepresent in S via robotic proxy, since this will preventH_1 being affected by the perceptual cues of H_2, which may hamper H_1's communicational performance relative to assessment questions (e.g. a female student may answer exam questions more freely if she cannot detect the gender of her examiner). Symmetric telepresence in our view will not produce what we call below the "Fair Proxy Effect", a specific phenomeno that arises in a human interlocutor when she or he is in the direct physical presence of the robotic proxy.

Comment 3: According to (Def1-2), the remote operation capabilities of the robotic proxy should preserve the communicative capacities of H_1 as these are relevant for D. For our illustration of the job interview, this means that the robotic proxy should be able to produce H_1's verbal and non-verbal input to the conversations in ways that do not diminish the semantic content of what is said; for example, H_1's voice should be transmitted in real-time without delays, sufficiently loud and clearly, in natural prosody, her head movements should be sufficiently expressive and accompany her speech in real-time without delay etc. In a different scenario, other communicative capacities of H_1 may be relevant – for example, if FPC is used to remove potential ethnic or racial biases in an audition for a choir, the focus will be on the precise rendition of all qualitative aspects of spontaneous vocal productions while head movements will be irrelevant. Or again, if FPC were used in an exam for choreographers where candidates are asked toimprovise a display of their choreographic

creativity, the functionalities of the robotic proxy would be focused on the accurate rendition of the bodily movements of H_1 dancing his or her improvisation in the remote location. The fact that (Def1-2) requires that the functionalities of the robotic proxy transmits *all* communicational elements that are relevant for D and *only* these, may create considerable technical difficulties (e.g. the servo motors of the robotic proxy must not be audible etc.).

Comment 4: As the formulation of (Def1-3) conveys, FPC may be used to remove different sorts of perceptual biases, i.e. the positive or negative valuations that members of a society connect with different perceptual stereotypes and that may be guided by in their decision-making. FPC may be used to remove any or all of the perceptual cues that provide information about H_1's gender, ethnicity, race, age or social class. Since the practical-ethical purpose of FPC is to increase *perceived fairness* in a communicational scenario of type S, i.e. to increase the fairness of D relative to the knowledge of H_1 and H_2 about implicit bias. We assume here that H_1 and H_2's knowledge is representative of the state-of-the-art of research on implicit biases in a society at a given time. The goal in (Def1-3) is called 'practical-ethical' since it is a practical goal (removal of perceptual cues for bias) pursued for the sake of an ethical value (fairness). Obviously, it is an underlying assumption in (Def1-3) that H_2 is not informed beforehand about H_1's gender, age, race etc. that is, the feature of H_1's social identity that is to be masked by way of using FPC.

Comment 5: It is constitutive for FPC that the reduction of perceptual bias is primarily due to the use of the robotic proxy. We consider a robot as a bundle for affordances (in the Gibsonian sense) – they afford perceptual and practical interactions by humans and many of their physical and kinematic features afford perceptual interactions relative to social categories. These affordances seem to vary across cultures and across individuals, but extensive research in Human-Robot Interaction studies is necessary to clarify the extent of this variation. By the ‘design’ of a robot, we mean its affordances constituted by its three-dimensional visual appearance, its kinematic, acoustic, phonetic, tactile and even olfactory features. To construct a robotic proxy is a difficult task since not only must the proxy fail to have affordances that are the cues for certain perceptual biases (e.g. neither its appearance, voice or movement must give the gender away), it must also retain the affordances of 'smooth' direct dialogue with a fellow human being.

With (Def1) and these five supplementary comments, we hope to have delineated the concept of FPC sufficiently precisely, yet also sufficiently generally, in order to convey that FPC is special communicational format that has a wide scope for practical applications while raising a host of research questions, as the following two sections will elaborate. To faciliate the exposition in these sections, we close this section by introducing two further concepts. The second definition is a mere abbreviatory stipulation:

Definition (Def-2) A *fair proxy scenario* is a communicative situation for which all conditions of (D1) are fulfilled, i.e. a scenario where FPC is de facto taking place or can take place.

The third definition, however, is an attempt to capture a distinctive phenomenology (i.e., distinctive experience that subjects are or can make themselves aware of by introspection) that participants reported in qualitative interviews accompanying past and ongoing pilot and experimental studies. The communicative scenarios of these studies are counselling scenarios (conflict mediation and ethical counselling) and we used Telenoid™ (model R1 and R4) robots as robotic proxies, as shown in Fig. 3.

Figure 3. doi

The Telenoid robot created by Hiroshi Ishiguro, ATR Hiroshi Ishiguro Lab, Japan.

The creator of the Telenoid, Hiroship Ishiguro, describes the design idea of the Telenoid robots as follows. The Telenoid "was designed to appear and to behave as a minimalistic human; at the very first glance, one can easily recognise the Telenoid™ as a human while the Telenoid™ appears as both male and female, as both old and young. By this minimal design, the Telenoid™ allows people to feel as if an acquaintance in the distance is next to you. [The Telenoid is] like an empty screen on to which specific features of the remote conversation partner can be projected" (Ishiguro 2012). It is questionable, however, whether the Telenoid's design indeed evokes the phenomenology that Ishiguro postulated (Yamazaki et al. 2012, Leeson 2017); at least in the context of Fair Proxy scenarios, the envisaged projection of features of a specific human being did not occur or was at least not in the foreground of what was reported. Participants who conversed with a counsellor via the Telenoid (i.e. who were then in the role of H_2) did not report that they were 'filling in the screen' and imaginatively supplemented the missing features that would convert the perceptual impression of the "minimalistic" human being into the impression of a normal human being with full-blown specific information. Instead, they:

- reported the experience of being in the presence of a strangely indeterminate or generic human being and
- reported cognitive relief.

Participants attributed the experienced cognitive relief to the fact that the missing social cues made it impossible to second-guess and anticipate appraisal by the interlocutor–"it was a robot so one would not need to think about one's own behavior"; "I could allow myself more [to be me]"; talking to the counsellor via the Telenoid "made it easier to

concentrate on what was being said", to "control [their] emotions better" and to "think better".

These reports suggest that the initially implicitly perceived absence of perceptual cues about social identities (gender, age, race, ethnicity etc.) was something that subjects eventually (after 1-2 minutes of getting used to the unfamiliar situation) were in some fashion aware of or could make themselves be aware of upon subsequent questioning and experienced as in the emotional context of relief, which seems to them connected to the absence of the normal procedures of social epistemic alignment with a dialogue partner or self-censoring.

A careful and extensive investigation of this phenomenology is currently being undertaken, but in order to facilitate the formulation of research hypotheses, we wish to present the following tentative definition:

Definition (Def-3): In a Fair Proxy scenario, human subjects in the role of H_2 (i.e. who interact with the robotic proxy in bodily proximity) experience a distinctive phenomenology or subjective impression that consists of three elements:

1. they experience that they are in the direct presence of a human being;
2. they experience the absence of information pertaining to social identities;
3. this absence of specific information is experienced in connection with the emotion of relief and greater focus and freedom to concentrate on the task at hand.

This distinctive subjective impression does not arise when FPC is not used in the given type of communicative scenario. We call this distinctive phenomenology or subjective impression the *Fair Proxy effect*.

The degree of variation in the Fair Proxy effect across different Fair Proxy scenarios is still an open question. In particular, it is still an open question whether element 3 of the Fair Proxy effect, which we so far have identified in scenarios of (conflict and ethical) counselling, can also be observed in assessment communication. Moreover, it is an open question whether the Fair Proxy effect is unique in interacting with the Telenoid or will also arise when robotic proxies are implemented with robots of different design.

Practical significance

The two top values that drive the research, design and development process for the robotics application we call Fair Proxy Communication are social justice and peace. These two values demarcate the main areas of concrete practical use of FPC as we currently envisage it.

Social justice is violated when certain members of society are discriminated against, i.e. deselected or devalued in an assessment procedure based on features that are irrelevant for the assessment in question. Discrimination often occurs at the level of implicit social cognition, when pre-conscious stereotyping or perceptual biases guide the conscious

assessment or decision in the context of job interviews, hiring or promotion. A large body of research suggests that, despite anti-discrimination laws, candidates with overt characteristics that are associated with negative interviewer evaluations (such as gender, ethnicity, being pregnant, overweight or LGBTQ) are frequently discriminated against (Macan and Merritt 2011, Heilman and Eagly 2008, Duguet et al. 2015), based on implicit perceptual biases (Bertrand et al. 2005, Agerström and Rooth 2011, Sekaquaptewa et al. 2003, Ziegert and Hanges 2005). The ideal – but humanly impossible – situation for a fair selection process would be one where the candidate is:

1. personally present in direct view,
2. without camouflaging distortions of the human shape, in order to enable the natural state and fluency of real time direct dialogue in 3D and yet
3. does not offer perceptual cues to personal or social identity.

Robots, on the other hand, here can do "what humans should but cannot" – they can help people to be bodily present without also presenting perceptual cues of their social identities.

The use of FPC for the sake of reducing (perceived) discrimination and increasing (perceived) social justice thus is in full compliance with the non-replacement maxim (see section "Background" above) and at least *prima facie* an area where social robotics can be responsibly employed. We envisage that FPC will be of practical interest for:

- assessment communication with equal opportunity employers (job interviews, promotions etc.)
- public educational institutions holding oral final exams
- public institutions holding any type of hearing (court hearings) where direct dialogical settings are important but potentially compromised by discrimination

Perceptual biases in business and educational communication (job interviews, wage negotiation, exams) create high losses. Gender and race discrimination not only offend principles of social equality endorsed by many countries, it also lowers an economy's total output — according to recent studies, a 50% increase in the gender wage-gap lowers the country's GDP up to 15-25% and the global net loss is calculated to amount to several trillion US dollars (Jacobsen 2011).

Before FPC can be taken into use, however, further extensive research is necessary to better understand precisely *how* it should be used. A core question here is whether and how any increase in *perceived* social justice or reduction of *perceived* discrimination translates into actual decisionmaking and persistent cultural change (see next section).

The second main application area which should benefit from the use of FPC is the mediation or facilitation of conflicts – here an increase in perceived justice of the communicative situation can increase opportunities for interpersonal peace (with possible extensions to intergroup peace). If the mediator / facilitator is represented by the robotic proxy, this will prove beneficial in conflict where gender, ethnicity, race or age play a central role, for example, especially in divorce conflicts, where the gender of the mediator /

facilitator often has negative effects on the mediation (see e.g. Jehn et al. 2010, Poitras 2009, Pradel et al. 2006). So-called perceived mediator neutrality is also an important factor in ethnic conflicts. Such conflicts are often intractable due to the fact that a mediator who has intimate familiarity with the context and its history most likely also has an ethnicity of one of the parties. Several European studies show that the monetary gain from early conflict resolution would be profound (see e.g. De Paolo et al. 2011).

Besides assessment communication and conflict resolution, there is surely a wide variety of other possible uses for FPC. In general, one could consider applying FPC in communicative contexts where direct communication of one party should be combined with anonymity of the other party; this can be useful for communication contexts where there is an asymmetry of power, such as complaint hearings between staff and management or between students and teachers etc.

Research tasks

Fair Proxy Communication and the possibility of communicating via robotic proxy via telepresence in general, raise a host of new and highly interdisciplinary research questions involving robotics, neuropsychology (cognitive science), psychology, gender research, business and management studies, conflict research, communication science, linguistics, anthroplogy, education science and philosophy (ethics). Here we shall only set out some of these research tasks with focus on one discipline, but it will be quickly apparent that the relevant research questions reach beyond the discipline and required interdisciplinary collaborations. (To abbreviate or clarify the exposition, we shall use the five variables we introduced in (Def1), S standing for the communicational scenario; D standing for the decision to be taken on the basis of S; H_1 standing for the person represented by the robotic proxy; H_2 standing for the person interacting directly with the robotic proxy; and PKD standing for the physical and kinematic design of the robotic proxy.)

As neuroscientists and neuropsychologists have begun to notice, social robots can be used as a new sort of research instrument to study social cognition (see e.g. Oberman and Ramachandran 2007, Wiese et al. 2017, Wykowska et al. 2015). Social robots offer a unique opportunity systematically and independently to vary the values of all those parameters that may be relevant for certain aspects of social cognition – e.g. degree of linguistic mirroring, gaze following, synchronisation of type and speed of body language, degree of human-like shape and voice, degree of gender, degree of epistemic alignment, degree of expression of emotions etc.; moreover, they can be used to create verifiable and reliably repeatable data variations; finally, they allow for direct interaction in 3D physical space, which increases ecological validity.

The phenomena of FPC offer a particularly well-focused entry point into this new research field of neuropsychological social cognition research with robots.

1. *The explanation of the Fair Proxy effect*: A first research task for neuropsychology would seem to try to explain the peculiar phenomenology of the Fair Proxy effect

(D-3 above). At first sight, it would seem plausible that the absence of triggers for the complex processes of implicit social cognition should result in the experience of greater focus – if the cognitive system is unburdened of the tasks of epistemic alignment and self-censoring, the task at hand (e.g. evaluation of H_1's performance, search for solutions to an emotional or ethical problem) can be more freely explored. For example, can the Fair Proxy effect be related to neurological indicators that were postulated as explanations for the influence of external motivation (anticipation of being judged by someone else) on the regulation of stereotypes responses (in the context of racial bias) (see e.g. Amodio et al. 2006, Amodio and Swencionis 2018)?

2. *Modulating implicit social cognition*: A much more complex and general research task arises with the question of how should we design fair proxies for different scenarios S. A robotic proxy is a complex affordance structure carried by its design PKD. How closely do alterations of PKD correspond to alterations in the processes of implicit social cognition? And can we design the affordances presented via PKD in such a way that those are left out that pertain to perceptual biases as relevant in S, yet retain those affordances that are necessary so that H_1 can exercise their communicative capacities and fulfil the functions that H_2 is to evaluate with respect to decision D? (For example, if the PKD modulates the voice to mask gender or even removes dialectical deformations to mask regional origins, how will this affect H_2's impressions of competency, which often hinge on timing and natural ability - fast answer, delivered with ease?) The very idea that, via different PKDs, we can create certain affordance structures that split the complex processes of implicit social cognition like a prism, so that FPC filters out those affordances that may negatively affect just decision-making in S, rests on the assumption that the processes of implicit social cognition are sufficiently modular and separable. This assumption in itself needs to be investigated. (For example, a 'gender-neutral' voice cannot be achieved by merely modulating pitch – other aspects like prosody and semantics play into our perception of gender). Furthermore, the relative significance of perceptual cues varies from context to context and perhaps also over time. (For example, while the voice of robotic proxies in our counselling scenarios carried litte significance for gender attributions, the acoustic affordances will receive much more attention if S is screening for a position as radio host. Moreover, currently it is unclear whether the comparatively small effect of the voice as cue for gender will not be increased as FPCs become more common). Thus FPC also offers an opportunity to investigate the dynamics of perceptual cues over time – of exposure or number of trials. In short, FPC provides an excellent entry point into the exploration of implicit social cognition by systematic interventions in physical and kinematic affordances structures – without robots, these kinds of systematic interventions are not possible. Another large complex of research tasks arises in the area of conflict and negotiation research. Here the third element of the Fair Proxy effect is of particular interest.
3. *FPC and cooperative rationality*: To the extent that a conflict or negotiation may be negatively affected by the implicit processing of social identity cues such gender, ethnicity and race, FPC may influence the decision-making of the parties involved

in a variety of ways, raising far-reaching research questions in the areas of conflict and negotiation research, psychology and anthropology. Will FPC make it easier to resolve a conflict or to reach an agreement? Will there be differences in the type of interaction patterns or neogtiation styles leading to a conflict resolution or an agreement? If so, on both questions, which of the phenomenon elements of the Fair Proxy effect are in the foreground for the parties involved and thus can be most plausibly connected with these behavioural changes? Is it the lack of social identity cues of the mediator or facilitator that promotes constructive solutions or agreements, the fact that the mediator's or facilitator's counsel cannot be suspected of bias and thus weighs more? Or is it rather that parties in a conflict or negotiation become more creative once processes of self-censorship are disabled, once parties experience the cognitive relief of being 'permitted' to concentrate on the task at hand without second-guessing? Or are parties more inclined to cooperate when all the information they receive from their dialogue partner is conspecificity – being in the presence of a fellow human being? Another complex set of research questions pertains to the role of technology as "the fourth mediator" (Vestergaard) in these mediation or facilitation settings. The novelty of the communicative situation, i.e. the creativity required by people in the role of H_2 in particular, may engender a 'frame of mind' or special cognitive regime that may be conducive for cooperative behaviour, at least relative to certain personality types.

4. *FPC and the elimination of bias and discrimination in assessment communication*: FPC has been cautiously defined with respect to *perceived* potential for biased or just decisions. A central set of research tasks pertain to the overall question of how well FPC performs in the praxis of assessment communication – in a globalised world with rising migration but also resurgent ethnocentrism pushed by populist political parties, little could be more important for the maintentance of social peace than efficient instruments against discrimination, either in the form of training programmes or technology or the combination of both. So much empirical research in gender studies, social psychology, sociology, business and management studies, but also education science, is needed in order to clarify the effectiveness of FPC as a tool for anti-discrimination vis-a-vis and in combination with other means currently in use, such as public awareness raising or training programmes. There are three levels of effectiveness of FPC to be distinguished: (1) the short-term phenomenological changes in H_2 we labelled the Fair-Proxy effect (Def-3 above); (2) the behavioural effect in H_2's decision-making – which decision is taken but also how, i.e. comparative consistency and time spent on decision-making; (3) the long-term effect of the use FPC towards the dissolving of perceptual biases and stereotyping, once the results of FPC are communally reflected in the context of a company's or institution's explicit commitments to equal opportunity. A large-scale research project would be necessary to investigate the relationship of these three levels of effectiveness for different types of communicative scenarios of assessment.
5. *FPC and the ethics of nudging*: There are certain forms of ethical enhancement that are themselves questionable from an ethical point of view – depending on which general position in normative ethics one adheres to (e.g. consequentialism,

deontology, virtue ethics, care ethics), the idea of promoting fairer decisions by manipulating a person's implicit, i.e. preconscious, perceptual 'mechanisms' can appear highly problematic. FPC thus also engenders a new line of debate in the area of "roboethics" and the debate about ethical enhancement. Does FPC amount to an objectionable form of ethical "nudging"? Even if it is not as invasive as ethical enhancement by medication, an important research task will be to define, informed by the empirical research results from the preceding research tasks, precisely how FPC should be embedded into other phases of the overall communication scenario or should be followed up by activities of individual and communal reflection (within a company or institution) to ascertain that those deliberative processes ensue, relative to which one can maintain the goal pursued by FPC, is indeed an ethical value in a sense that all standard varieties of normative ethics can agree to.

Implementation and acknowledgements

A research team of 26 researchers from 14 disciplines currently investigated the phenomena of FPC, its foundations and applications, in the context of a larger research project supported by the Carlsberg Foundation (2016-2021). Following the method of Integrative Social Robotics, we investigated FPC in wide-scope interdisciplinary interactions. For an introduction to the overall idea of the project, see (Seibt 2016a); an overview over the researchers involved can be found at the project website www.integrativerobotics.org.

Acknowledgements

The research announced in this project is supported by a Semper Ardens grant of the Carlsberg Foundation. We thank our colleagues in the project group for productive interactions that benefited the exposition of the concept of FPC as formulated here. The presentation in the section "Research Tasks" includes ideas from (in alphabetical order) Lin Adrian, David Amadio, Malene Flensborg Damholdt, Dan Druckman, Charles Ess, Michael Filzmoser, Cathrine Hasse, Sabine Köszegi, Marco Nørskov, Sladjana Nørskov, Josh Skewes and John Parm Ulhøi.

Funding program

Carlsberg Semper Ardens Grant Program.

Grant title

What Social Robotcs Can and Should Do–Towards Integrative Social Robotics.

Hosting institution

Aarhus University, Denmark

References

- Agerström J, Rooth DO (2011) The role of automatic obesity stereotypes in real hiring discrimination. Journal of Applied Psychology 96: 790-80. https://doi.org/10.1037/a0021594
- Amodio DM, Kubota JT, Harmon-Jones E, Devine PG (2006) Alternative mechanisms for regulating racial responses according to internal vs external cues. Social Cognitive and Affective Neuroscience 1 (1): 26-36. https://doi.org/10.1093/scan/nsl002
- Amodio DM, Swencionis JK (2018) Proactive control of implicit bias: A theoretical model and implications for behavior change. Journal of Personality and Social Psychology 115 (2): 255-275. https://doi.org/10.1037/pspi0000128
- Bertrand M, Chugh D, Sendhil Mullainathan (2005) Implicit Discrimination. American Economic Review 95 (2005): 94-98. https://doi.org/10.1257/000282805774670365
- De Paolo G, Feasley A, Orecchini F (2011) Quantifying the Cost of Not Using mediation – a Data Analysis. European Parliament, Brussels. http://www.europarl.europa.eu/document/activities/cont/201105/20110518ATT19592/20110518ATT19592EN.pdf
- Duguet E, Parquet LD, L'horty Y, Petit P (2015) New Evidence of ethnic and gender discriminations in the French labor market using experimental data: A ranking extension of responses from correspondence tests. Annals of Economics and Statistics/Annales d'Économie et de Statistique 2015: 21-39. https://doi.org/10.15609/annaeconstat2009.117-118.21
- Dumouchel P, Damiano L (2017) Living with Robots. Harvard University Press [ISBN 9780674971738] https://doi.org/10.4159/9780674982840
- Friedman B, Kahn P, Borning A (1997) Value sensitive design and information systems. In: Zhang P, Galetta D (Eds) Human-Computer Interaction in Management Information Systems. Routledge, New York.
- Heilman ME, Eagly AH (2008) Gender stereotypes are alive, well, and busy producing workplace discrimination. Industrial and Organizational Psychology 1: 393-398. https://doi.org/10.1111/j.1754-9434.2008.00072.x
- Ishiguro H (2012) The Telenoid Robot. http://www.geminoid.jp/projects/kibans/Telenoid-overview.html. Accessed on: 2017-11-10.
- Jacobsen J (2011) Gender Inequality. A Key Global Challenge: Reducing Losses due to Gender Inequality. https://www.copenhagenconsensus.com/sites/default/files/gender.pdf . Accessed on: 2017-1-11.

- Jehn KA, Rupert J, Nauta A, Van Den Bossche S (2010) Crooked conflicts: The effects of conflict asymmetry in mediation. Negotiation and Conflict Management Research 3 (4): 338-357. https://doi.org/10.1111/j.1750-4716.2010.00064.x
- Leeson C (2017) Anthropomorphic Robots on the Move: A Transformative Trajectory from Japan to Danish Healthcare. PhD Dissertation University of Copenhagen
- Macan T, Merritt S (2011) Actions speak too: Uncovering possible implicit and explicit discrimination in the employment interview process. International Review of Industrial and Organizational Psychology 26: 293-337.
- Nourbakhsh I (2013) Robot Futures. MIT Press [ISBN 9780262018623]
- Oberman LM, Ramachandran VS (2007) The simulating social mind: the role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. Psychological Bulletin 133 (2): 310-327. https://doi.org/10.1037/0033-2909.133.2.310
- Poitras J (2009) What makes parties trust mediators? Negotiation Journal 25 (3): 307-325. https://doi.org/10.1111/j.1571-9979.2009.00228.x
- Pradel D, Bowles H, McGinn K (2006) When Gender Changes the Negotiation. Harvard Business School, Working Knowledge URL: http://hbswk.hbs.edu/item/5207.html
- Seibt J (2016a) Integrative Social Robotics–Towards Culturally Sustainable Technology Solutions. http://www.carlsbergfondet.dk/en/Forskningsaktiviteter/Forskningsprojekter/Semper-Ardens-forskningsprojekter/Johanna-Seibt_Integrative-Social-Robotics. Accessed on: 2016-7-01.
- Seibt J (2016b) Integrative Social Robotics—A New Method Paradigm to Solve the Description Problem and the Regulation Problem? What Social Robots Can and Should Do—Proceedings of Robophilosophy/TRANSOR. IOS Press, Amsterdam, 10 pp.
- Seibt J, Damholdt M, Vestergaard C (2018) Five Principles of Integrative Social Robotics. In: Coeckelbergh M, Loh J, Funk M, Seibt J, Nørskov M (Eds) Envisioning Robots in Society—Power, Politics, and Public Space. IOS Press, Amsterdam.
- Sekaquaptewa D, Espinoza P, Thompson M, Vargas P, von Hippel W (2003) Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. Journal of Experimental Social Psychology 39: 75-82. https://doi.org/10.1016/S0022-1031(02)00512-7
- Turkle S (2011) Alone Together. Basic Books
- Van den Hoven J (2005) Design for values and values for design. Information Age 4: 4-7.
- Wiese E, Metta G, Wykowska A (2017) Robots as intentional agents: Using neuroscientific methods to make robots appear more social. Frontiers in Psychology 8: 1663. https://doi.org/10.3389/fpsyg.2017.01663
- Wykowska A, Chaminade T, Cheng G (2015) Embodied artificial agents for understanding human social cognition. Philosophical Transactions of The Royal Society B Biological Sciences 371: 20150375. https://doi.org/10.1098/rstb.2015.0375
- Yamazaki R, Nishio S, Ogawa K, Ishigur H (2012) Teleoperated android as an embodied communication medium: A case study with demented elderlies in a care facility. 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication. IEEE, Paris, 1066–1071 pp. https://doi.org/10.1109/ROMAN.2012.6343890

- Ziegert J, Hanges P (2005) Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. Journal of Applied Psychology 90: 553. https://doi.org/10.1037/0021-9010.90.3.553