New approaches to the emerging social neuroscience of human-robot interaction

Anna Henschel (BSc, MSc)

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

October 2020

School of Psychology
University of Glasgow
62 Hillhead Street
Glasgow
G12 8QB

Abstract

Prehistoric art, like the Venus of Willendorf sculpture, shows that we have always looked for ways to distil fundamental human characteristics and capture them in physically embodied representations of the self. Recently, this undertaking has gained new momentum through the introduction of robots that resemble humans in their shape and their behaviour. These social robots are envisioned to take on important roles: alleviate loneliness, support vulnerable children and serve as helpful companions for the elderly. However, to date, few commercially available social robots are living up to these expectations. Given their importance for an ever older and more socially isolated society, rigorous research at the intersection of psychology, social neuroscience and human-robot interaction is needed to determine to which extent mechanisms active during human-human interaction can be co-opted when we encounter social robots.

This thesis takes an anthropocentric approach to answering the question how socially motivated we are to interact with humanoid robots. Across three empirical and one theoretical chapter, I use self-report, behavioural and neural measures relevant to the study of interactions with robots to address this question. With the Social Motivation Theory of Autism as a point of departure, the first empirical chapter (Chapter 3) investigates the relevance of interpersonal synchrony for human-robot interaction. This chapter reports a null effect: participants did not find a robot that synchronised its movement with them on a drawing task more likeable, nor were they more motivated to ask it more questions in a semi-structured interaction scenario. As this chapter heavily relies on self-report as a main outcome measure, Chapter 4 addresses this limitation by adapting an established behavioural paradigm for the study of human-robot interaction. This chapter shows that a failure to conceptually extend an effect in the field of social attentional capture calls for a different approach when seeking to adapt paradigms for HRI.

Chapter 5 serves as a moment of reflection on the current state-of-the-art research at the intersection of neuroscience and human-robot interaction. Here, I argue that the future of HRI research will rely on interaction studies with mobile brain imaging systems (like functional near-infrared spectroscopy) that

allow data collection during embodied encounters with social robots. However, going forward, the field should slowly and carefully move outside of the lab and into real situations with robots. As the previous chapters have established, well-known effects have to be replicated before they are implemented for robots, and before they are taken out of the lab, into real life. The final empirical chapter (Chapter 6), takes the first step of this proposed slow approach: in addition to establishing the detection rate of a mobile fNIRS system in comparison to fMRI, this chapter contributes a novel way to digitising optode positions by means of photogrammetry.

In the final chapter of this thesis, I highlight the main lessons learned conducting studies with social robots. I propose an updated roadmap which takes into account the problems raised in this thesis and emphasise the importance of incorporating more open science practices going forward. Various tools that emerged out of the open science movement will be invaluable for researchers working on this exciting, interdisciplinary endeavour.

Table of Contents

List of Tables

List of Figures

Dedication

I dedicate this thesis to my family.

“To understand ourselves as a species is one of the profound undertakings of a lifetime.”

- Kahn and colleagues (2006)

Acknowledgements

The Greek scholar Archimedes supposedly exclaimed “Heureka” (‘I have found it!’) when he stepped into a bath and realised that the water level rose proportionate to his body immersed under water. And while I conducted studies in the pursuit of this joy of scientific discovery, I now want to take the opportunity to exclaim the Glaswegian “Yaldi!”, which expresses excitement (and in my reading of the term) deep gratitude for the great minds that supported me along the way.

When I started my PhD, I was swept away by my supervisor Prof. Emily Cross’s infectious enthusiasm and remember like it was yesterday when I waited nervously for my first meeting to discuss ideas for the many scientific projects I envisioned. I will be ever grateful that I had the freedom to go where my intellectual curiosity led me, even though I sometimes encountered some dead ends along the way. Emily was always there to cheer me on, support me and helped me reframe any ‘challenges’ as ‘opportunities’.

I started my PhD in a small Welsh town and was lucky enough to encounter the most welcoming and friendly postgraduate community there. I am so thankful to my former flatmates Tom, Rikesh and Beckie, who participated in every single one of my pilot experiments and who organised Halloween, Valentine’s day and dance parties to simply celebrate a Tuesday evening.

When I moved to Glasgow, I was lucky again: Dr. Ruud Hortensius joined my supervisory team. Throughout my PhD, he taught me to not only focus on doing good science, but also to keep a positive mental attitude™ and treasure a healthy work-life balance. I am forever indebted to him for sharing his wisdom, and especially his joy for science. Academia is a better place because of scientists like him. I would also like to thank my third supervisor Prof. Lisa DeBruine, who in the final months of the PhD stepped in to guide me in moments of doubt.

I also want to take a moment to acknowledge the role models who crossed my path, and who inspired me to pursue a PhD in the first place; Prof. Belinda

Campos at University of California Irvine, PD Dr. Astrid Rossegger at Konstanz University, Prof. Valeria Gazzola and my former supervisor Dr. Selene Gallo at the Netherlands Institute for Neuroscience. Through these strong, female role models I felt empowered to pursue this path myself.

As my supervisor has written elsewhere, science is a cumulative and a collaborative endeavour, and none of this work would have been possible without the constant support and encouragement of the famous #TeamSoBots. I especially want to thank robot-whisperer Bishaka Chaudhury, who helped me learn to speak the language of our robotic friends, and for teaching me many important life lessons, as well as being the best colleague and office mate one could wish for.

I am also grateful for the support I received from the Glasgow admin staff, who went above and beyond to facilitate my research endeavours, and especially Fiona Dick, who showed me what real Glaswegian hospitality was by opening her home to me when I temporarily needed a place to stay. Not only did I receive a room to stay in, but also flowers on my bedside table, an emotional support cat and treats to take to work. Thank you for everything you have done for me.

It will be difficult to find a community that was as supportive as the Glasgow Institute of Neuroscience and Psychology. I want to highlight Prof. Niamh Stack's invaluable mentorship, the encouragement by the #PsyRuns team and the many fun discussions with Dr. Maria Gardani at our Glasgow Reading Minds Book Club. I am so grateful for the support of Dr. Elizabeth Adams, who encouraged many of my creative and professional endeavours. I would like to thank Stephanie Allan who co-organised the Glasgow ReproducibiliTea journal club with me and became a great friend. Thank you to the PhD and open science communities, online and offline, who have been there through the ups and downs of this journey.

When I left the safety of the Glasgow nest, I was greeted with so much kindness: thank you to Prof. Paulo Boggio at Mackenzie Presbyterian University in Sao Paulo, to Prof. Peter Keller and the whole MARCS team at Western Sydney University for their warm hospitality.

I am grateful towards my brilliant research assistants: Hannah Bargel, who was my first intern and whose dedicated research skills inspired me, Lauren Colbert, Luca Leisten, Julia Fechner and Rosanne Timmerman. It will be exciting to see where this talented and thoughtful next generation of scientists is going.

Thank you, friends, who in the final months of my PhD showed unwavering support, cooked for me, sent gifts, motivational letters and helped me design the cover for the print version of this thesis. Thank you especially to Nina Kattler, Alexandra Eyfjord and Laura Immler for absolutely everything.

All the writers, thinkers, musicians and artists who made my life much richer and have kept me company in one form or another over the last three years – thank you for sharing your creative gift with the world. In “A Room of One’s Own”, Virginia Woolf reflects on women’s place in literary history, a domain traditionally dominated by men. She proposes that “a woman must have money and a room of her own if she is to write” (p.4). In writing this doctoral thesis, I was privileged that my family provided me with both, gave me the tools and the space to pursue my scientific passion, even under the most difficult of circumstances. Beyond giving me these resources, they did so much more. I especially want to thank my father for driving me to a remote German supermarket in the middle of a global pandemic, so I could record new stimulus material. I would like to thank my mother for taking every single one of my emotional phone calls and for sending me countless care packages. And I want to thank my sister, my pillar and my inspiration, who spent hours on Zoom just to keep me company when I struggled with particularly challenging manuscript revisions.

In her essay, Virginia Woolf also writes of the 17th century aristocrat Lady Margaret Cavendish, whose work, she argues, suffered because of a lack of accessible education. Woolf writes:

> “She should have had a microscope put in her hand. She should have been taught to look at the stars and reason scientifically. Her wits were turned with solitude and freedom.” (p.71)

I want to thank my family, for literally putting microscopes in my hand as a child, telling me stories about the stars, and for nurturing my scientific hunger.

Author’s declaration

Research Output

Related to this thesis

Published output

Chapter 3 - Henschel, A., & Cross, E. S. (2020). No evidence for enhanced likeability and social motivation towards robots after synchrony experience. *Interaction Studies*, 21(1), 7-23. https://doi.org/10.1075/is.19004.hen

Chapter 5 - Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social Cognition in the Age of Human-Robot Interaction. *Trends in Neurosciences*. https://doi.org/10.1016/j.tins.2020.03.013

Under review

Chapter 4 - Henschel, A., Bargel, H., & Cross, E. S. (2020). Faces do not attract more attention than non-social distractors in the Stroop task. Accepted pending minor revision at *Collabra: Psychology*. https://psyarxiv.com/pbfny

Chapter 1 – Henschel, A., Laban, G., Cross, E. S. (2020). Charting the Rise of Social Robots, from Fiction to Hype to Reality. Invited manuscript submitted for publication at *Current Robotics Reports*.

Other work

Under review

Pownall, M., Talbot, C., Henschel, A., Lautarescu, A., Lloyd, K. E., Hartmann, H., Darda, K. M., Tang, K. T. Y, Carmichael-Murphy, P., & Siegel, J. (2020). Navigating Open Science as Early Career Feminist Researchers. Invited submission currently under review at *Psychology of Women Quarterly*. https://psyarxiv.com/f9m47

Contributorship

In the following, contribution summaries are listed for each of the thesis chapters. For one chapter, a more detailed contributor statement is available in the Contributor Roles Taxonomy (CRediT) format, this is marked with an asterisk.

Chapter 1

AH conceptualized and wrote the chapter; ESC critically reviewed and edited the chapter.

Note: This chapter formed the basis for an invited manuscript currently submitted to *Current Robotics Reports* (see Henschel, Laban & Cross, under Research Output).

Chapter 2

AH conceptualized and wrote the draft, AH and ESC critically reviewed and edited it.

Chapter 3

AH conceptualized the chapter; BC programmed the task; AH, VD, LL, AG, and AA collected the data; AH carried out statistical analyses and wrote the manuscript; AH and ESC critically reviewed and edited the manuscript. AH and ESC approved the final version for publication.

Chapter 4

AH, ESC, GT and KK contributed to conceptualisation and design; AH, HB and LC collected the data, AH, HB, ESC, AM and ES contributed to the analysis and interpretation of the data; AH wrote the manuscript; AH, HB, and ESC reviewed and edited it; AH, HB and ESC approved the final version for publication.

Chapter 5

AH, RH and ESC conceptualised, wrote, critically reviewed and edited the manuscript; AH, RH and ESC approved the final version for publication.

Chapter 6*

AH and RH conceptualised and designed the experiment; AH, RH, LML and HS carried out data collection; AH, RH, MK, PP, KS and LML contributed to the analysis and interpretation of the data; AH and MK prepared the draft; AH, RH, ESC, MK and PP critically reviewed and edited the draft.

Chapter 7

AH conceptualized and wrote the draft, AH and ESC critically reviewed and edited it.

Key

AH: Anna Henschel; ESC: Emily S. Cross; RH: Ruud Hortensius; BC: Bishakha Chaudhury; VD: Veronica Diveica; LL: Leorah Lasrado; AG: Alex Gillespie; AA: Abdulla Almusalam; GT: Guillaume Thierry; KK: Kami Koldewyn; HB: Hannah Bargel; LC: Lauren Colbert; AM: Andrew Milne; ES: Eline Smit; LML: Luca Marie Leisten; HS: Hanna Seelemeyer; MK: Michaela Kent; PP: Paola Pinti; KS: Katharina Stute

Chapter 1 Introduction

In the beloved children's classic 'The Adventures of Pinocchio', a wooden marionette comes to life through the hands of master puppeteer Geppetto (Collodi, 1883). Pinocchio ventures out into the world and after a long series of misfortunes eventually learns to show empathy and kindness towards Geppetto and others. The puppet's understanding of fundamental human virtues finally culminates in its metamorphosis into a real boy. The story of this transformation, from a lifeless object to a fully functioning human, has been re-told in various forms in the science fiction literature, and echoes a principal desire in robotics today: to create a machine that perfectly embodies the traits that make us fundamentally human and fundamentally social - a robot that is independent and can learn.

Since its inception, the scientific field of robotics has been closely intertwined with science fiction literature, with the first mention of the word robot made by Karel Čapek in his 1920 play 'Rossum's Universal Robots' (Hockstein et al., 2007). In this play, robots who look almost indistinguishable from humans are exploited as factory slaves and later rebel against their human makers, another popular trope in science fiction. A bit later, the term 'robotics' was coined by Isaac Asimov, in his short story 'Liar!', which features a robot that is compelled to lie so as not to upset its human creators (1941). While these terms were introduced historically quite late, visions of automata have existed for almost as long as humans have lived together in societies. From ancient Egypt, Greece and China, to the 18th century 'Turk' (a fake chess playing machine, which in fact was controlled by a human hiding inside the device) and the friendly Japanese 'Gakutensoku' - mechatronic puppets and automatons have fuelled the public imagination of what might be possible in terms of human-fabricated autonomous agents that interact with us - almost as equals (Frumer, 2020; Schwartz, 2019).

Goodrich and Schultz (2007) remark in their survey on human-robot interaction (HRI) that the impact of science fiction literature on robotics cannot be denied: the inventors of the very first industrial robot – 'Unimate', a mechanical arm deployed at the General Motors car factory in the spring of 1961, were initially inspired by Asimov's stories. The authors reflect on the relatively young field of human-robot interaction, which they define as “the field of study dedicated to understanding, designing and evaluating robotic systems for use by or with humans” (p. 204). In the late 1960s a big leap forward was accomplished when Nilsson published his work on the first autonomous robot (‘Shakey’), which was able to navigate around a block obstacle course (Kuipers et al., 2017; Nilsson, 1969).

These breakthroughs in autonomous robotics, as Broadbent (2017) and others have argued, were facilitated by developments in the nascent field of artificial intelligence and the foundations laid by Alan Turing in his work on digital computing (1950). Goodrich and Schultz (2007) cite the first meeting of the IEEE Symposium on Robot & Human Interactive Communication (RoMan) in 1992 and the first explicitly multidisciplinary meeting of the ACM International Conference on Human-Robot Interaction (ACM-HRI) in 2006 as defining moments in the emergence of HRI as a scientific field of study. In addition to these initial scientific meetings, the authors mention engineering and technical challenges as another important catalyst, such as the RoboCup Search and Rescue competition (Goodrich & Schultz, 2007). Further, the authors note that the field of human-robot interaction has been driven by its applications, and they reflect that major developments have been facilitated in part due to interest in domains such as search and rescue, and space exploration. This is also evident in the ‘three Ds of robotization’: robots for dangerous, dull and dirty work (Takayama et al., 2008).

Early enthusiasm for the potential of robotics is perhaps best illustrated with Bill Gates’ essay in the Scientific American (2008) ‘A robot in every home’, which envisioned that in the near future, robots would become part of our everyday lives – much like the personal computer. These robots, he wrote, would help with various tasks in the household, and in addition to providing assistance, would also provide companionship. While his vision of ubiquity has not quite come true yet, it is the case that many modern households employ robotic vacuum cleaners, like the Roomba robot, or speech-based personal assistants, like the Alexa system (Šabanović, 2010; Vallverdú & Trovato, 2016), suggesting his vision might be slowly but surely moving toward reality.

1.1. Defining social robots

Within the field of human-robot interaction, social robots take on a special role, and fall under the category of 'proximate interaction', in which "humans and robots interact as peers or companions" (Goodrich & Schultz, 2007, p. 205). In a bibliometric analysis by Mejia and Kajikawa (2017), it becomes apparent that the social robotics literature comprises only a small part of the larger robotics knowledgebase: 2.3%, to be exact. According to their search in the Web of Science database, the authors identified discussions on social robotics appearing as early as 1970, but, as the authors illustrated, it was not until the late 1990s that the field started growing rapidly. Based on reference information of the extracted articles, the authors were also able to identify relevant clusters that represent the social robotics knowledgebase. The largest clusters in social robotics research can be summarized as 'robots as social partners' and 'human factors in human-robot interaction'. Interestingly, Mejia and Kajikawa (2017) also point out that research trends emphasize the various fields of application for social robots: robots as companions, robots as educators for children, and robots as assistants for the elderly. This is consistent with a trend identified by Šabanović, who in interviews with robotics researchers in the US and Japan identified that social robots "often represent technological fixes", i.e. using a technological approach to solve a pressing societal problem (2010, p. 349). Furthermore, when investigating what constitutes the majority of the knowledgebase in social robotics, Mejia and Kajikawa (2017) find that even though they play a central role, the social sciences *are hardly represented* (Figure 1). The authors write aptly: "Social robotics is social in its intention, but its knowledgebase is concentrated in the engineering and technology domains" (p.11).

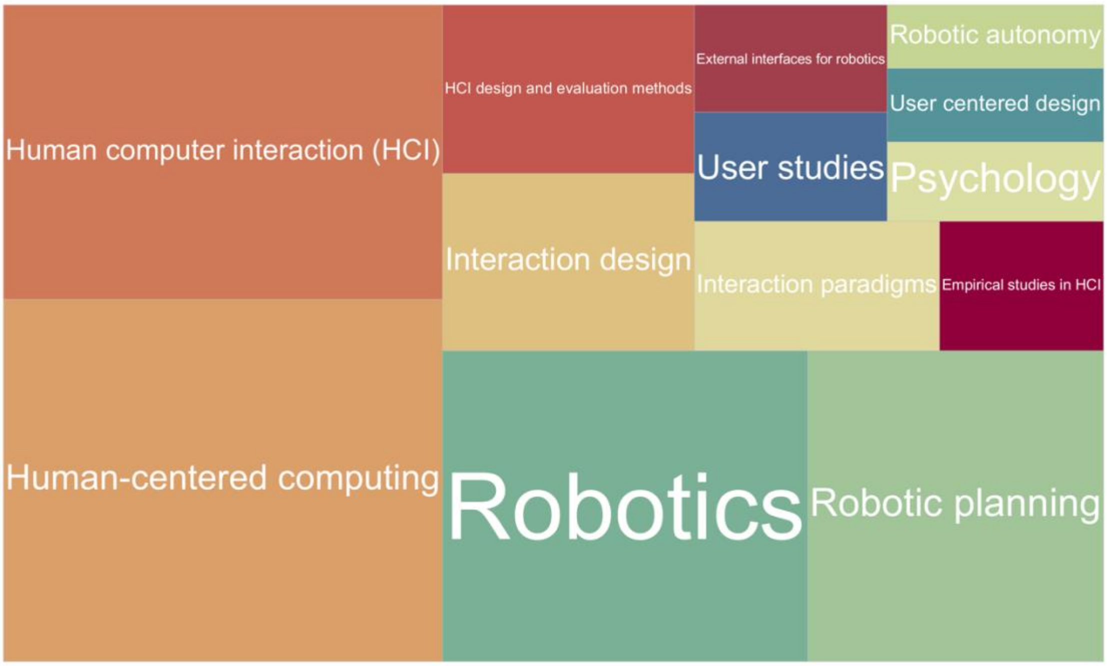


Figure 1 - Subject areas in the ACM-HRI conference proceedings

The subject areas are presented as a tree map with the size of the area representing the number of conference proceedings in each category. Robotics being one of the most prominent categories (947 results), there are some nods to the afore-mentioned social sciences: psychology (143 search results) and user studies (175 results). Data taken from: https://dl.acm.org/conference/hri

Indeed, while the interdisciplinary nature of social robotics is emphasized throughout the literature, this observation by Mejia and Kajikawa reveals an interesting tension that has also been voiced by Broadbent (2017) and Eyssel (2017) – the literature could benefit from knowledge about the mechanisms of human social behaviour gained in psychology, the cognitive science and neuroscience (which are here referred to as ‘social sciences’, but depending on the country and higher education conventions are occasionally considered as part of science, technology, engineering, or mathematics). This issue is discussed in more detail in Chapter 2.

When reviewing the social robotics literature, it becomes apparent that there is not a generally agreed upon understanding of what social robots are, and what effectively constitutes a robot as being ‘social’ continues to be negotiated and debated by various authors. Sarrica and colleagues (2019) investigated the question of how social robots are understood by analysing definitions in articles published by the International Journal of Social Robotics between 2009 and 2015. An overview of the most popular definitions they identified is presented in

Table 1. By investigating the most often cited definitions, it becomes apparent how heterogenous the understanding of social robots is.

Despite this lack of homogeneity, Sarrica and colleagues (2019) were able to identify a few shared traits of the definitions: social robots are physically embodied agents that have some (or full) autonomy and engage in social interactions with humans, by communicating, cooperating and making decisions. These behaviours are then interpreted by human onlookers as 'social', according to current norms and conventions. It is of note that in discussions of what constitutes a social robot, many authors listed in Table 1 acknowledge that a truly social robot, as described in their definition, remains a vision of the future. Lee and colleagues (2006) emphasize that we are still far away from sophisticated social robots depicted in popular movies like Stephen Spielberg's A.I. Artificial Intelligence and Dautenhahn (2007) remarks that she remains sceptical: "It is unclear whether the 'social-emotional' dimension in human-human interaction can be fulfilled by robots, whether the inherently mechanical nature of HRIs can be replaced by truly meaningful social exchanges" (p. 701). Interestingly, this point is somewhat at odds with her argumentation that exploring social competencies for robots might actually be the missing piece in building stronger artificial intelligence.

Table 1 - Popular definitions of social robots in the literature, identified by Sarrica and colleagues (2019).

Authors	Year	Key term(s)	Definition
Breazeal	2003	sociable	"Denoting robots that pro-actively engage with humans, having their own internal goals and needs in order to satisfy internal social aims (drives, emotions, etc.). These robots require deep models of social cognition not only in terms of perception but also of human modelling." (p. 169)

Fong	2003	socially interactive	“We describe robots that exhibit the following ‘human social’ characteristics: express and/or perceive emotions; communicate with high-level dialogue; learn models of or recognize other agents; establish and/or maintain social relationships; use natural cues (gaze, gestures, etc.); exhibit distinctive personality and character; and may learn and/or develop social competencies.” (p. 145)
Duffy	2003		“A physical entity embodied in a complex, dynamic, and social environment sufficiently empowered to behave in a manner conducive to its own goals and those of its community.” (p. 177)
Bartneck & Forlizzi	2004		“A social robot is an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioural norms expected by the people with whom the robot is intended to interact.” (p.592)
Lee	2006		“Social robots are a new type of robot whose major purpose is to interact with humans in socially meaningful ways.” (p.962)
Dautenhahn	2007	socially intelligent	“A robot companion is a robot that (i) makes itself ‘useful’, i.e. is able to carry out a variety of tasks in order to assist humans, e.g. in a domestic home environment, and (ii) behaves socially, i.e. possesses social skills in order to be able to interact with people in a socially acceptable manner.” (p. 685)
Hegel	2009	social interface (form,	“A social robot is a robot plus a social interface. A social interface is a metaphor which includes all social attributes by which an observer

	function, context)	judges the robot as a social interaction partner.” (p. 174)

A few authors go further and present comprehensive frameworks. For example, Kahn and colleagues (2006) argue that social robots should be evaluated against ‘psychological benchmarks’, which capture fundamental aspects of human life. Using illustrative scenarios of what our future lives with social robots could look like, the authors raise important questions about whether it is in the best interest of companies and stakeholders to produce fully autonomous robots. Giving a benign example, the authors ask: would users want a robot to disagree with them about which card game to play? Might it be problematic to think of social agents as ‘useable’ objects? Who should be held accountable, and do these robots possess intrinsic moral value? In total nine possible benchmarks are described, which illustrate the central problem Kahn and colleagues, as well as many other social robotics researchers, characterize: “to understand ourselves as a species is one of the profound undertakings of a lifetime” (p. 384). This point is further elaborated in Chapter 2, where I illustrate the importance of social robots for gaining more understanding about the human brain and social behaviour.

Finally, Baraka and colleagues (2019), in the face of the growing diverse landscape of social robots, propose their extended framework, by illustrating seven relevant dimensions of social robots: their appearance, the social capabilities, autonomy and intelligence of the robot, the proximity and temporal profile of the interaction and the context of the interaction (i.e., its purpose and application area). In their appearance classification system, they distinguish between bio-inspired robots (these can be human- or animal-inspired), artifact shaped (for example robots resembling man-made objects or those that are imaginary) and functional robots (for example drones).

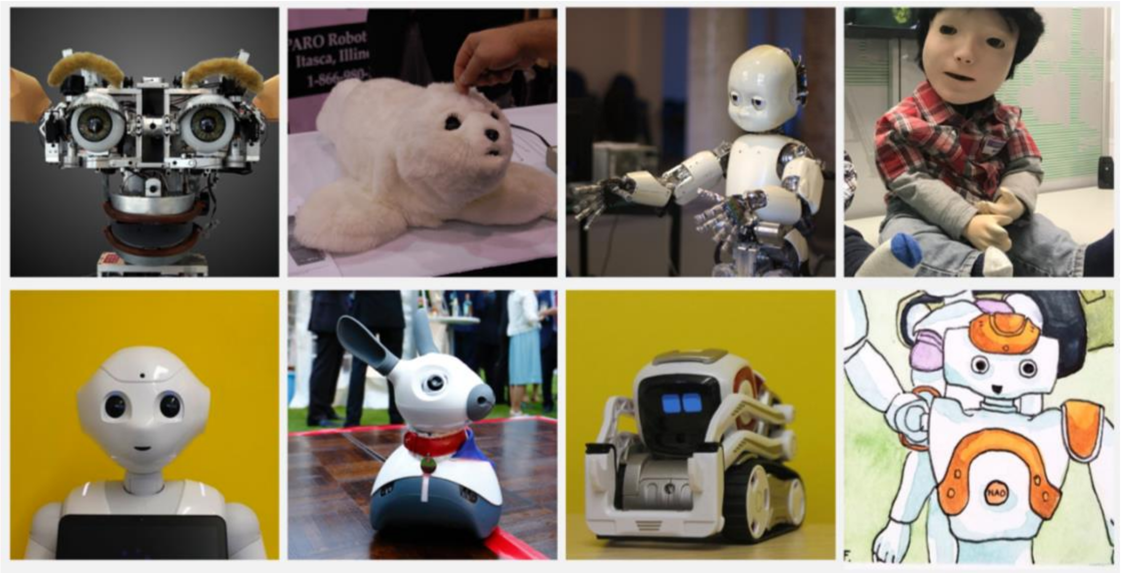


Figure 2 - Robots introduced in this chapter (and mentioned throughout this thesis)

First row, from left to right: Kismet (1997), Paro (2003), iCub (2004) and the Kaspar (2006) robot. Second row, from left to right: Pepper (2014), MiRo (2015), Cozmo (2016) and Nao (2008). The ABOT database (http://www.abotdatabase.info/) and the IEEE "ROBOTS" website (https://robots.ieee.org/) provide more comprehensive overviews of available platforms. Image sources: Kismet by Rama on Wikipedia (CC BY-SA 3.0 fr); Paro Therapy Robot by Theron Trowbridge on Flickr (CC BY-NC 2.0); iCub by Jiuguang Wang on Flickr (CC BY-SA 2.0), Kaspar by Loz Pycock on flickr (CC BY-SA 2.0);); https://www.so-bots.com/robots; MiRo by UK in Japan- FCO on Flickr (CC BY 2.0). Nao artwork by Julia Fechner.

After this initial ontological exploration, I will now briefly consider social robotics through the lens of popular platforms for research and their application areas outside of experimental research.

1.2. Overview of popular social robots in scientific research

In the following, robotic platforms are introduced which have fundamentally shaped the field of social robotics (Figure 2). This list is not intended to be representative of all commercially or custom designed robots that are currently available. Instead, it is intended to highlight some of the main social robots that are currently in use, the different applications that are being envisioned for social robots, and the various avenues of research that are currently being pursued.

1.2.1. Kismet

Many consider Kismet to be the first ‘intelligent’ social robot (Breazeal & Scassellati, 1999). Kismet was designed following the principle of building a sociable robot (Table 1), which, as the authors reason, relies on in-depth models of social cognition (Breazeal, 2003). Kismet has rudimentary facial features that resemble those of a human and can convey positive and negative facial emotional expressions by dynamically moving its eyelids, eyebrows, mouth, and ears. Breazeal and Scassellati (1999) modelled Kismet’s behaviour on that of young infants, whose mothers, despite limited means of explicit communication, interpret infant behaviour as intentional and infant proto-typical speech as meaningful. Kismet is programmed to look for perceptual cues in the environment that satisfy its internal drives for socialness, stimulation or rest. By means of implementing this software architecture, the authors argue that they have implemented proto-social responses that will convey intentionality of the robot towards its human partners. Following an initial proof-of-concept evaluation (Breazeal & Scassellati, 1999), human-robot interaction experiments with Kismet have laid the foundation for the design of social robots that are able to detect and mirror human emotions (Breazeal, 2003).

1.2.2. Paro

This social robot, designed to look and behave like a baby harp seal, was developed by Shibata and colleagues at the National Institute of Advanced Industrial Science and Technology in Japan (Shibata et al., 2003). The Paro robot can move its eyelids, head, front and back flippers, listens via microphone and produces harp seal sounds via speakers. It responds to tactile stimulation by means of sensors located at various points across its body and thus interactively responds to the user’s touch. After initial favourable evaluations in science museums in Japan and the UK pertaining to the robots perceived likeability, Paro has been deployed and evaluated in the context of elderly care homes as a therapeutic companion for patients with dementia (Shibata et al., 2003). As one of the most successful and most-widely sold social robots, several studies have shown that though its behavioural repertoire is very limited, its interactive capabilities and appearance have had positive effects on elderly participants and patients suffering from dementia (Broadbent, 2017; Kidd et al., 2006; Robinson et al., 2013; Selma et al., 2013). Data from a randomized controlled trial and a long-term observational study indicated that interactions with the seal robot reduced loneliness and increased opportunities for interactions among the care home residents (Kidd et al., 2006; Robinson et al., 2013).

1.2.3. iCub

Much like Kismet, iCub, the ‘robot child’, is based on theories of developmental psychology and cognitive neuroscience (L. Natale et al., 2017; Sandini et al., 2004). Formerly known as RobotCub, this robot was developed as a testbed for the theory of embodied cognition. This theory describes the phenomenon of learning and development via physical interaction with the world through a human(oid) body (Sandini et al., 2004). Like a child exploring its environment, iCub was designed to manipulate its surroundings, imitate its human partners and communicate with them. iCub has been used in cognitive neuroscience studies to investigate whether humans perceive it as intentional and as an agent with a mind (Ghiglino et al., 2020; J. Perez-Osorio et al., 2018). Across several studies, it has been shown that especially knowledge cues about the behaviour of the robot influence the degree to which participants perceived the robot’s intentionality (Wiese et al., 2017; Wykowska et al., 2016).

1.2.4. Kaspar

Another field of application for social robots that is currently being explored is the context of social skills training for children with an autism spectrum condition (ASC). Kose-Bagci and colleagues (2009) argue that the reduced behavioural repertoire of the Kaspar robot might offer an opportunity to interact with children who lack social interaction skills typical for their age and their level of development, hallmark indicators of an ASC. Here, the humanoid form of the robot is preferred, as the authors speculate that the social skills learning may translate to human-human interactions outside the lab. However, the evidence base for this line of research is considered weak, and additional studies have repeatedly been invited to investigate these questions with larger sample sizes (Broadbent, 2017; Pennisi et al., 2016).

1.2.5. Pepper

The commercially produced humanoid Pepper robot (SoftBank Robotics) has been used in a multitude of human-robot interaction studies to investigate its social acceptance, its role in educational contexts and general attitudes towards the robot (Jacobs, 2018; Kennedy et al., 2013; Thunberg et al., 2017). The robot was originally built as a business-to-business product for SoftBank Robotics stores in Japan, however, it eventually transformed into a popular platform to attract customers into stores worldwide (see Figure 3), and became a tool in academic research (Pandey & Gelin, 2018). Its design was informed by evaluations of the previous model, the Nao robot (Figure 2). Users expected robots to be taller than the 58cm-tall Nao, so Pepper's height was targeted at the height of a person sitting in a chair (Pandey & Gelin, 2018). Pepper's design also references themes from Japanese culture, including its manga-inspired eyes and the hip joint, which allows the robot to bow (Pandey & Gelin, 2018). This robotic platform was used in Chapter 3 of this thesis, and more details on its degrees-of-freedom and inbuilt features are given there. While Pepper is the next generation of SoftBank's robots, a study by Thunberg, Thellman and Ziemke (2017) found that participants were more likely to comply with the request of a Nao robot, compared to the Pepper robot's requests to return a book to the experimenters. Mubin and colleagues (2018) investigated its use in public spaces and found that it was less popular than the Nao robot (maybe owing to the robots' respective price points), with more papers designing and evaluating interaction scenarios for the Nao robot in public spaces. The authors identified papers that evaluated Pepper's social acceptability in a shopping mall, an elderly care home, in a remote classroom and as a customer service employee in a hotel lobby scenario (Aaltonen et al., 2017; Stock & Merkle, 2018; Tanaka et al., 2015; Yang et al., 2017).

1.2.6. MiRo

Taking a different approach, the MiRo robot is a biomimetic system, whose design does not aim to imitate human social cognition, but rather the brain and behaviour of a simpler mammalian animal (Collins et al., 2015). The developers explicitly justify their design choice of animal morphology as a strategy to mitigate potential disappointment of users and their expectations towards the social capabilities of the robot. The design of the robot features light patterns under the translucent shell of the back, which satisfies two goals: the simple communication of affect and increasing the salience of the interaction with an artificial, rather than a real, social agent (Collins et al., 2015). The robot, which evokes a pet-like impression, includes characteristics taken from “puppies, kittens and rabbits” (Collins et al., 2015, p. 2). The robot is described as an edutainment product, which alludes to its intended purpose as an educational tool for children. However, the robot has also been explored as a fall alert system, relevant especially to the population of over 80-year-olds, who are expected to fall in their home environments at least twice a year (Georgiou et al., 2020). In their proof-of-principle study, the authors demonstrate that MiRo could be used as a mobile and smart tool to locate a person on the ground, and send a help signal if no movement of the person is detected.

1.2.7. Cozmo

This commercial robot was initially developed by Anki (USA) as a children’s toy. However, the research community has embraced the palm-sized Cozmo (Figure 2), especially for its engaging behavioural animations and expressive, emotional facial expressions, which were informed by an animator who worked on ‘Wall-E’, the Disney Pixar movie (Chaudhury et al., 2020; Skågeby, 2018). Initial studies show that long-term interactions in a socializing intervention with this social robot do not enhance human-like empathic responses towards it when it is shown in simulated pain, and another recent study implies that while participants behave reciprocally towards the robot in an economic game, there is no evidence for pro-social behaviour towards it (Cross, Riddoch, et al., 2019; Hsieh et al., 2020). As a platform for researchers, especially when it comes to long-term human-robot interaction, it could prove useful, as it collects a wide range of information about the user, similar to the behavioural sampling method (Chaudhury et al., 2020; Henschel, Hortensius, et al., 2020b). However, it should be noted that like many social robots, future research efforts might be limited by the demise of the company, which ultimately leads to ceased support for the robot (Hoffman, 2019). This common problem will be discussed in the following sections.

1.3. The social robot paradox

Duffy coined the 'social robot paradox', which has remained a critical point in social robotics over the years (2005, p. 1):

> "In fact, humanoid robots outside of science fiction, have thus far only been toys or research platforms with nebulous applications. It is intriguing that one of the most powerful paradigms for adaptivity and flexibility, the human, has so far, when modelled in the form of a machine, resulted in little more than a toy. Its usefulness is very limited."

16 years later, this observation still rings true, with new, commercial- or bespoke social robots (like Cozmo and MiRo) moving away from referencing the human form. Humanoid robots, like Pepper, are designed to be reminiscent of human characteristics, and at the same time avoid imitating every aspect of appearance to pre-empt an Uncanny Valley effect (Pandey & Gelin, 2018). This effect references the phenomenon that *almost* fully humanlike androids elicit eery or uncanny feelings in the user (Mori et al., 2012). While the humanoid shape as a design feature is a powerful signal to users that the agent affords social interactions, it also makes the robot more prone to failing to deliver on high expectations regarding the nature of the interaction (Dereshev et al., 2019; Kahn et al., 2006).

To investigate this phenomenon more closely, Dereshev and colleagues (2019) interviewed long-term, expert users of the Pepper robot. Their participants had lived and interacted with the robot on timescales ranging between 8 months and more than 3 years. The researchers report that one specific expectation regarding the humanoid Pepper robot was its ability to engage in a reciprocal conversation. The participants' expectations were disappointed when the robot was not able to go beyond the smart-speaker like single-turn structure of conversation. One of the participants also pointed out that people who interacted with Pepper quickly lost interest, a finding which is echoed in a usability study by Aldebaran (later purchased by SoftBank Robotics), where Pepper was deployed to the homes of users over several weeks (Rivoire & Lim, 2016). The novelty effect is a common problem in social robotics, and long-term studies have often found a reduced engagement with various robotic platforms over time (Leite et al., 2013; Tanaka et al., 2007).

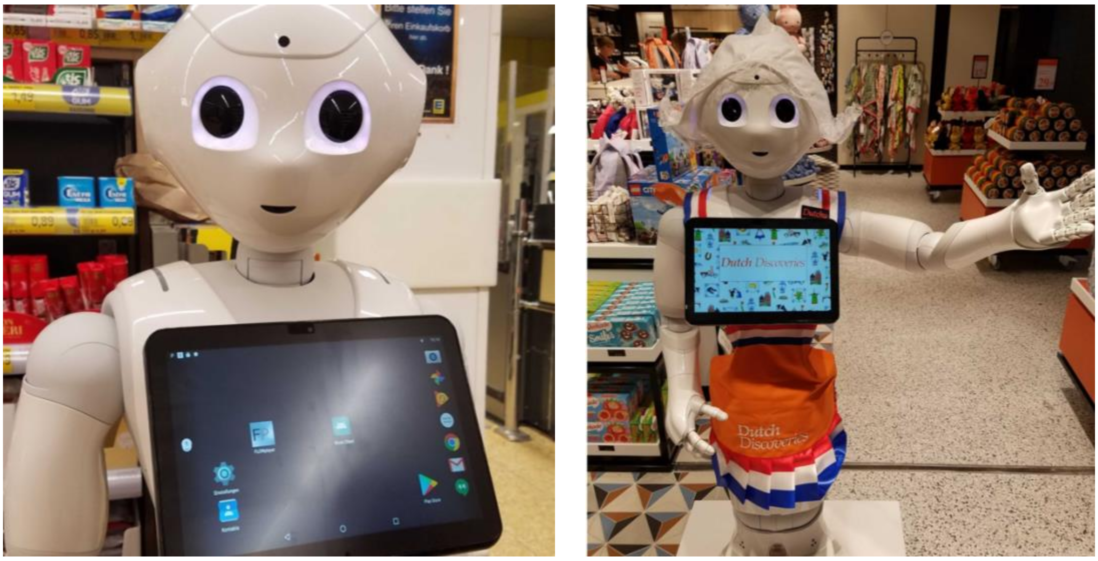


Figure 3 - Examples of the Pepper robot 'in the wild'

Left: The social robot was placed at the customer checkout in a German supermarket and reminded shoppers of the new hygiene regulations in April of 2020, during the global coronavirus pandemic. Right: Another Pepper robot in a Dutch souvenir shop at Schiphol airport. Pictures taken by the author.

Thus, the robots we are familiar with through science fiction films like 'Ex Machina' or 'Robot & Frank' remain a futuristic dream, whether they are depicted as helpers and companions, or villains (Broadbent, 2017; Garland, 2014; Schreier, 2012; Wiese et al., 2017). When we encounter robots 'in the wild' (Figure 3), this discrepancy between the reality of social robots and our expectations towards them becomes even more salient. The Pepper robot, when used in supermarkets or airports, becomes little more than a puppet repeating the same script over and over, as the autonomous capacities of the robot are not advanced enough to sustain natural interactions in unconstrained, changing environments, especially when groups of people are present. Social interactions are a complex problem, to humans and robots alike. Designing and programming robots to successfully and seamlessly integrate into the human social world remains an especially "wicked problem" (Hannibal & Weiss, 2020, p. 1).

Duffy wrote about these issues in the early 2000s, observing that machines are simply not to be equated with humans, and thus every attempt to design socially accepted machines may only result in a colourless impression of key human social characteristics and traits (Duffy, 2004). Since then, researchers have echoed this sentiment when they reflect on the commercial failures of many social robotics start-up companies and the fact that as of 2019, there are still few fully-autonomous commercial robots available to consumers (Dereshev et al., 2019; Goodrich & Schultz, 2007; Hoffman, 2019; Tulli et al., 2019; Wiese et al., 2017; Yang et al., 2018).

1.4. Towards truly sociable robots?

Currently available social robots remain puppets in the hands of private consumers or research teams, not yet having realised their potential to becoming fully autonomous and transforming into sophisticated social agents. When encountered outside of the lab, these social robots often execute scripts (with puppet masters/programmers pulling the strings) and have little interactive capacities.

In the late 1990s, Breazeal and Scassellati entitled their work on Kismet “robots that make friends and influence people” – however, to this day, this appears still utopic, with serious implications for the many fields of applications in which social robots are expected to prompt innovation right now.

When thinking about the future with truly sociable robots, one can perhaps turn to an unexpected avenue for inspiration: animal models of social collaboration (Collins, 2019; Kahn et al., 2006). In a seminal study by Halloy and colleagues (2007), the researchers took advantage of self-organising behaviour of social animal societies, like the cockroach. The authors developed behavioural models for robotic cockroaches (that looked nothing like their biological counterparts), placing emphasis on appropriate behavioural responses and the correct chemical signal, which they determined were the key factors to the acceptance of robots in a cockroach group. Indeed, these robots autonomously interacted with the animals, and drove collective decision making, demonstrating social acceptance in the group and collaboration. Circling back to human-robot interaction, one key factor in designing successful interactions with humanoid robots will therefore be to “maximise […] both its mechanical advantages and [find] the minimum humanlike aspects required for their social acceptance” (Duffy, 2006, p. 33).

1.5. Summary

In this chapter, I have briefly introduced the field of social robotics, by reflecting on its early developments in the 1990s, as well as crucial advances in autonomous robotics that have led to an increased interest and research and development investment into the field. Social robotics is a small subdiscipline of human-robot interaction that envisions robots as assistants and companions. As this introduction highlights, it is a heterogenous and multidisciplinary field. Important questions raised by researchers early on concern the ethical implications of developing fully autonomous, truly social robots. However, in the field’s relatively short history, many scientists acknowledge the limitations of currently available social robots. Indeed, by reviewing the most successful robotic applications, it becomes apparent that despite their potential to be used as companions for the elderly, as educators, or teachers for children with ASC, many questions remain regarding the capabilities of robots to take on more social roles, especially if they are to be working alongside human users autonomously. Studies by Dautenhahn and colleagues show that participants in their studies still do not see robots as companions or friends, but rather as useful household servants (2007). Despite this, the research reviewed in the context of these robotic platforms hints at their usefulness as a testbed for human social cognition, in terms of probing its flexibility and dimensions (Hortensius & Cross, 2018a).

Chapter 2 An interdisciplinary approach to investigating questions in social cognition

2. The role of the humanoid robot in experimental research – implications for experimental rigor and reproducibility

The African Ubuntu philosophy states that personhood emerges through interactions with others (Birhane, 2017; Bolis & Schilbach, 2018). In their recent philosophical perspective on social cognition, Bolis and Schilbach (2018) emphasize that this dynamic perspective on personhood puts the construction of the self through social interaction front and centre. Importantly, this process is not only initiated by other humans, but also when we change our environment ourselves. From this viewpoint, tools become a part of the exchange, and here the authors explicitly include artificial intelligence and robots. According to them, a dialectical view would likely yield more advanced artificially intelligent and social systems, as well as facilitate the search for answers to fundamental questions of social cognition. To that effect, this chapter presents an update to the philosophical tenets of Vygotsky (Bolis & Schilbach, 2018; Vygotsky, 1987): "We become ourselves through others – including robotic others"

By studying human-robot interaction and developing more advanced social robots, we are forced to reverse-engineer what it means to be human, what constitutes social interaction and what makes interpersonal relationships successful. On the other hand, experimental psychological might positively influence the future development of social robots (Wiese et al., 2017). As alluded to at the beginning of Chapter 1, the longstanding human fascination for trying to understand the human psyche better by means of creation is also a recurring theme in science fiction, inspiring gothic writers like Mary Shelley and Jewish folklore, where the golem is created from clay (Kieval, 1997).

While contemporary robotics still suffers from the social robot paradox, we now have a tool at our fingertips that constitutes a (near-)perfect stimulus and allows us to answer fascinating questions about the dynamic changes in the social fabric of our society. Humanoid robots are an ideal control group for studying interactions with other people, as they allow us to probe the limits and the flexibility of social cognition (Henschel, 2019; Hortensius & Cross, 2018a).

Bolis and Schilbach's (2018) perspective matches the tenets of those researching at the intersection of social robotics, experimental psychology and cognitive neuroscience. In her seminal review, Broadbent (2017) writes (p. 629):

> "The further we go down the path toward making and interacting with artificial humans, the more truths we learn about ourselves."

Indeed, integrating humanoid robots into empirical research is now an established and rich tradition in the multidisciplinary fields congregating around the social robot, including robotics, experimental and social psychology and cognitive neuroscience (Chaminade & Cheng, 2009; Chevalier et al., 2020; E. S. Cross, Hortensius, et al., 2018a; Kompatsiari et al., 2018; J. Perez-Osorio et al., 2018; Wiese et al., 2017; Wykowska et al., 2016). The most prevalent recurring theme is that robots afford researchers interested in studying social interactions the benefit of ecological validity. While traditional research in social neuroscience used screen-based images and videos of socially relevant cues, humanoid robots allow scientists to retain experimental control while studying how people engage with a physically present agent.

When including humanoid robots in embodied interaction experiments, the experimenter has control over the verbal and non-verbal cues of this agent and can take advantage of the social presence the robot exerts (Henschel, 2019; Kompatsiari et al., 2018; Wykowska et al., 2016). While some have argued that deploying humanoid social robots in the wild may lead to exaggerated expectations of the robot's abilities, the subtle cues to humanness of the humanoids in their design and behaviour are ideally suited to experimental research, where they are steered by experimenters (Kahn et al., 2006). In the HRI literature, this is also referred to as 'wizarding' the robot, making a nod to the beloved American children's classic novel 'The wonderful wizard of Oz' (Baum, 1900). In the story, a young girl is transported to a magical land, where she encounters the powerful wizard of Oz. However, the girl and her companions finally realise that the magic is all smoke and mirrors: the great and powerful wizard turns out to be a fraud hiding behind a screen.

Controlling the movements and the speech outputs of the robot in the lab allows experimenters to take on the role of the wizard, operating with a more socially advanced humanoid robot, presenting a form of illusory social ability that has sparked some concerns about deception (Broadbent, 2017; Prescott, 2017). Thus, the robots presented in the lab do not accurately reflect those met 'in the wild', however, even those robots that are wizarded by a human experimenter may not reach a level of socialness embodied by humans. Chevalier and colleagues (2020) acknowledge this limitation when they write that the humanoid robot might not elicit the same mechanisms that are at play when interacting with other humans, though these studies might nonetheless yield important insights in controlled lab environments.

A popular approach researchers use is to take effects or theoretical concepts from experimental psychology and cognitive neuroscience and adapt them for human-robot interaction experiments (Kahn et al., 2006; J. Perez-Osorio et al., 2018). Previous lines of research have especially focused on motor resonance (Chaminade & Cheng, 2009), empathy (Cross, Riddoch, et al., 2019; Rosenthal-Von Der Pütten et al., 2014), joint attention (Chevalier et al., 2020; Willemse & Wykowska, 2019) and adopting the intentional stance towards robots (Wiese et al., 2017). For example, Chaminade and Cheng (2009) argue that motor resonance plays a fundamental role in human-human interaction and conducted a series of studies to test how strongly the actions of a humanoid robot arm would interfere with the motions of participants in a classic motor interference paradigm. The authors found that compared to an industrial robot arm, the humanoid robot (whose motions were informed by human motion capture) elicited a change in the movement of human participants, depending on whether the robot was moving its arm congruently or incongruently (Chaminade & Cheng, 2009). These experiments show the success of adapting well-understood concepts for HRI and demonstrate on a behaviour and brain level that our responses towards humanoid robots may be akin to fellow humans (Wykowska et al., 2016).

On the other hand, some researchers have argued repeatedly that HRI as a discipline lacks the systematicity characteristic of experimental research and have stressed that psychological and neuroscientific methods add rigor to the investigation of interactions with robots (Kompatsiari et al., 2018; Wiese et al., 2017, p. 1). According to Wiese and colleagues, motion-tracking and brain imaging measures are needed to "make robots appear more social" – an effort directed towards making the wizard hiding behind the screen redundant. These more sophisticated social robots, as I argue in Chapter 3, may be advantageous in many social contexts where they are framed as companions, carers and coaches. For example, Wiese and colleagues (2017) highlight that the perceived intentionality of social robots will play a major role in successfully orchestrated feelings of connection towards it, as well as better collaboration in shared environments. The perceived intentionality of the robot can only be achieved, the authors explain, if they are able to elicit activity in the hub regions of human social cognition. Thus, using neurophysiological measurements in HRI will contribute to a future with more sophisticated social robots (Wiese et al., 2017).

However, this argument does not remain uncontested by other researchers, who see relying on only neurophysiological methods as one-sided, especially in investigations into the mind perception of robots (Kewenig, 2018). Kewenig and colleagues (2018) argue that the study of social robotics cannot only depend on these electrophysiological measures alone, but must take a more holistic approach, considering for example self-report and qualitative data observed in interactions with robots.

2.1. Integrating perspectives from social robotics, neuroscience & psychology

Here some tensions between the different fields involved become apparent: while social robotics relies on insights from psychology and neuroscience, the lines of communication appear broken, and insights generated at the intersections of these disciplines are often not fed back to the engineers and developers of social robots. Baxter and colleagues examine these tensions more closely in their review of three years of proceedings of the ACM-HRI conference (2016). The authors note that there is a strong need for establishing a shared language between different disciplines and that the field should move towards a common ground, such as a community 'FAQ' to help establish a stronger collaborative infrastructure. The main difference between HRI and what has been classed the *social sciences,* is their research output (papers versus the more fast-paced conference proceedings), different specialist terminology, fundamental theories and methodological approaches. Thus, the communication gap between fields which has been described here is perhaps not surprising.

Interestingly, while researchers in experimental psychology and cognitive neuroscience mainly lament a missing objectivity in social robotics, HRI researchers observe a need for a more formalised structure of psychology and neuroscience knowledge, which might in the future contribute to lowering the barriers of engaging with other fields (Baxter et al., 2016). The dialogue between disciplines is fraught, as Hannibal and Weiss (2020, p. 1) write in their introduction to a special issue on current challenges and new methodologies in social robotics. As introduced in the previous chapter, these authors note that deploying sophisticated autonomous social robots remains a multifaceted challenge. Before robots can successfully integrate into a human-centric social environment, these complex challenges have to be addressed, which can only be achieved if all disciplines engage in a mutually respectful dialogue, acknowledging their own strengths and limitations. Hannibal and Weiss (2020) also observe a lack of consensus when it comes to methods and theory, a notion which is again echoed by Irfan and colleagues (2018a, p. 13), who adopt a more negative tone by characterising interdisciplinary ties between social psychology and HRI as an "uneasy marriage". These authors reference and criticise the standard approach described by Kahn and colleagues (2006), by acknowledging that the fields of psychology and neuroscience have recently undergone a crisis leading to some unreliable findings and negatively impacting the trust HRI researcher subsequently express towards these adjacent fields. When non-replicable concepts are carried over from social psychology to social robotics, studies are built on shaky grounds, they argue. Here the researchers call for a change in perspective, including raising awareness about the file-drawer problem, and admonishing questionable research practices. This point will be discussed in more detail in the final chapter of this thesis, building on the recommendations of Irfan and colleagues (2018).

In a similarly critical perspective on the interdisciplinary ties in HRI, Eyssel (2017) places a strong emphasis on the need for taking a theory-driven approach when conducting studies with humanoid robots. This renewed interest in building upon strong theoretical foundations is in line with a recently published perspective that sees the replication crisis as not solely exacerbated by questionable research practices, but also by a lack of overarching theoretical frameworks (Muthukrishna & Henrich, 2019). Eyssel (2017) draws on examples from her own work on adapting insights gained from intergroup relations research for a social robotics context. Here, dehumanising tendencies of outgroups become the starting point of investigation with humanoid social robots (Eyssel, 2017).

While many researchers have raised critical concerns about the current state of the field, they also acknowledge the benefits of collaborating with multiple disciplines. Irfan and colleagues (2018) do not suggest researchers abandon established foundations in psychological research, but instead propose that researchers working in this field take advantage of a renewed emphasis placed on rigor in all those engaging in HRI research. This could also be beneficial when it comes to the problem of “putting the cart before the horse”, in terms of developing social robots first and then operating with the constraints of their emotional and behavioural capabilities (Vallverdú & Trovato, 2016, p. 325). Broadbent (2017) observes that more and more psychologists are getting involved in HRI research and Baxter and colleagues (2016) commend the richness that the multi-disciplinarity affords the research questions. They also argue that this diversity should be protected and nurtured. Finally, the pioneers of using humanoid robots in experimental research perhaps summarise this positive development best when they close with the notion that each field can mutually benefit and learn from each other (Chaminade & Cheng, 2009).

2.2. Examining social behaviour toward robots: Social Motivation Theory

Situating the empirical work of this thesis according to the taxonomy proposed by Baxter and colleagues (2016) and following the recommendation of Eyssel (2017), my dissertation work has taken a theory-driven and distinctly human-centric perspective to HRI research. Indeed, the improvement of a social robot per se is not a central goal of this work, rather, as Baxter and colleagues (2016) have expressed it, improvements to social robots could be a practical, future outcome. There are a multitude of theoretical models one can adopt as a point of departure when thinking about social robots. Two popular early theories are the Media Equation by Nass and colleagues (1996), originating from communication theory, and the "like me" framework embedded in developmental psychology (Meltzoff, 2007). Nass and Reeves (1996) established in their studies that people would apply the same social rules to computers used in their experiments as they did towards other humans. The researchers observed that participants used more positive adjectives when the subsequent evaluation was presented by a computer they had worked on, compared to a computer unrelated to the previous task. This, for example, has resulted in studies investigating the effects of social robots openly rejecting participants, finding that their self-esteem was lowered when a humanoid robot stated after a game of Connect-4 that it would not like to see them again (Nash et al., 2017).

Another popular theory by Meltzoff represents the idea that fundamentally all social cognition originates in the recognition of other agents that are "like me" (2007). Empirical results from studies with very young children show that self-other distinction takes place at a young age (Meltzoff, 2007). For example, when adults mimic babies, they are preferentially fixated when they act like the child. This finding was extended by the same authors in the field of HRI, where the gaze behaviour of 18-month old infants was recorded during the interaction with a small humanoid robot (Meltzoff et al., 2010). Those infants were more likely to follow the gaze of the humanoid robot if they had previously observed the experimenter engage socially with the robot, than those infants who had not observed this exchange.

Overall, these theories seem to converge on the idea that humans tend to mindlessly apply social rules and conventions, even when confronted with inanimate agents, like computers or humanoid robots (Broadbent, 2017). This resonates with one of the most fundamental studies conducted in this area, which found participants applying social scripts and conventions to abstract moving shapes (Heider & Simmel, 1944). Moving circles and triangles were described as "lovers in the two-dimensional world" and framed in a social chase situation that suggested participants engaged in anthropomorphizing behaviour, despite minimal cues to animacy being present (Heider & Simmel, 1944, p. 247). Anthropomorphising describes the tendency to ascribe human characteristics and traits to non-human entities (Epley et al., 2008).

With this background, one of the questions I ask in this dissertation is whether and how interactions with humanoid robots are situated in terms of motivation. I used Chevallier and colleagues' (2012) theory as a springboard: this framework considers social motivation as a fundamental human drive, affecting behaviour, biology and evolutionary selection. The authors propose the Social Motivation Theory of Autism, in which they argue that individuals with ASC show diminished social motivation, which may have far-reaching downstream consequences. On a conceptual level, typically developing individuals are thought to show a special attentional bias toward social stimuli, which has been honed over the course of human evolution, due to associated evolutionary fitness of being quickly able to differentiate between friends or foes. Chevallier and colleagues (2012) emphasize that across several cognitive science experiments, it has been repeatedly shown that human faces and bodies are attended to rapidly and consistently. A second pillar of the theoretical framework relates to the reward value of social interaction: here the authors propose that we are inherently motivated to engage in social interaction, and that we seek it out where possible. The final pillar relates to social maintaining, i.e. people's drive to be seen in a favourable light and to (unconsciously) engage in behaviours that are conducive to sustaining interactions. For example, spontaneous movement imitation of others is referenced as a type of 'social glue' that fosters rapport and facilitates human social collaboration.

This theory has sparked much research in HRI with obvious links to the prioritized field of research of engaging children with ASC in playful learning environments with social robots. However, this theory has also led to researchers asking fundamental questions on the motivational value of humanoid agents (Chaminade & Okka, 2013; Simut et al., 2016). This thesis takes a similarly fundamental approach, and asks across social orienting and reward, how humanoid robots rank in terms of their motivational value.

2.3. Summary and overview of the thesis

This thesis is split in three parts. The first part (Chapters 1 and 2) has introduced core concepts and theories relevant to the thesis. This is followed by the main body, which includes published and submitted empirical and theoretical work on the subject. The third and final part of the thesis is a general discussion of the work as a whole. Author contributions are signposted in the section of the same name (see the Table of Contents).

To summarise, the first two chapters have aimed to give a short introduction to the field of social robotics by examining what ‘socialness’ means in the context of robotics and by considering applications and research through the lens of popular robotic platforms. Chapter 1 also reflected on the current challenges and limitations associated with social robots, while Chapter 2 explored how social robots are currently used in experimental research, the resulting interdisciplinary tensions, and, finally, introduced the theoretical frame of reference for the present thesis, situating the research in terms of the social motivational value of humanoid robots.

Chapter 3, the first empirical study, applied a phenomenon previously established in psychology and neuroscience, replacing a human interaction partner with a robot. Here we investigated whether the beneficial effects of interpersonal synchrony on rapport hold in this new context. In this between-subjects study, we identified a gap in the availability of reliable behavioural measures of social motivation, which led to the empirical study described in Chapter 4.

In Chapter 4, again, we adapted and conceptually extended a measure from cognitive neuropsychology. It was our aim to build on the eye-contact effect, which describes the phenomenon of human social cues claiming our attentional resources even when we are explicitly told to ignore them. It was our goal to extend this task by adding new stimulus material in the form of robot faces and pareidolic (object) faces, to adequately control for facial features in non-interactive situations and compare how the demand on attention would be shaped by these humanoid robot faces, in comparison to the socially more salient human faces.

In Chapter 5, we reflect on the importance of applying cognitive neuroscience methods to social robotics, however, taking a graded and stepwise approach. In this opinion piece, my co-authors and I reflect on the replication crisis and its implications for conducting studies at the intersection of social robotics, psychology and neuroscience. It becomes clear that these interdisciplinary studies should aim to carefully replicate well-known effects in psychology and neuroscience, before extending and adapting them for HRI, using humanoid social robots.

In the final empirical chapter (Chapter 6), I follow this proposed graded approach by identifying a brain region that plays a crucial role in governing human social interaction. In this study, I aimed to validate a new brain imaging system that allows mobile brain imaging in human-robot encounters, by identifying how brain activity recorded with functional magnetic resonance imaging (fMRI) compared to functional near-infrared spectroscopy (fNIRS). I used a robust localiser task, designed to elicit activity in the temporo-parietal junction. The motivation for testing the fNIRS system is its promise to further contribute to the ecological validity of experiments with humanoid robots by recording brain activity in natural interaction scenarios.

In the final chapter of the thesis (Chapter 7), I reflect on the results of the empirical and theoretical work (chapters 3, 4, 5 and 6), and place an emphasis on methodological considerations going forward. I provide a road map for future studies, highlighting the need for direct replication studies, new theoretical developments following up on Social Motivation Theory and suggestions for more transparent data visualisation.

Chapter 3 No evidence for enhanced likeability and social motivation towards robots after synchrony experience

This chapter is an exact copy of the author accepted manuscript of:

Henschel, A., & Cross, E. S. (2020). No evidence for enhanced likeability and social motivation towards robots after synchrony experience. *Interaction Studies*, 21(1), 7-23.

3. Abstract

A wealth of social psychology studies suggests that moving in synchrony with another person can positively influence their likeability and prosocial behavior towards them. Recently, human-robot interaction (HRI) researchers have started to develop real-time, adaptive synchronous movement algorithms for social robots. However, little is known how socially beneficial synchronous movements with a robot actually are. We predicted that moving in synchrony with a robot would improve its likeability and participants' social motivation towards the robot, as measured by the number of questions asked during a free interaction period. Using a between-subjects design, we implemented the synchrony manipulation via a drawing task. Contrary to predictions, we found no evidence that participants who moved in synchrony with the robot rated it as more likeable or asked it more questions. By including validated behavioral and neural measures, future studies can generate a better and more objective estimation of synchrony's effects on rapport with social robots.

3.1. Introduction

In his book *Deep Thinking*, former chess grandmaster Gary Kasparov (2017) recounts the story of his failure against the IBM super-computer Deep Blue in 1997. Contrary to what one might expect, he emphasizes that the triumph of the machine is ultimately the triumph of its human makers, and in order to thrive, humans must learn to live together with intelligent machines. Beyond chess playing devices, disembodied algorithms, and fully automatized factory lines, the present time is very much shaped by the rise of the *social* robots. These robots have the potential to provide society with economical care, company and therapy (Eriksson et al., 2005; Prescott et al., 2012; Robins et al., 2005). While robots are now deployed in various social contexts where they are framed as companions rather than tools (Darling, 2015; Duffy, 2000), roboticists and stakeholders are faced with the seemingly impossible challenge of making robots “truly social” (Duffy et al., 1999). Researchers describe this as a grand challenge with a vast problem space (Riek, 2014; Sandini et al., 2018). However, by endowing an artificial agent with socialness, patients as well as healthy individuals might benefit greatly from improved learning, companionship and therapeutic outcomes (Fasola & Matarić, 2012; Feil-Seifer & Matarić, 2011).

Wiese and colleagues (2017) suggest that the best way to make robots appear more social is to use the toolbox provided by neurocognitive research methods to implement empirically supported behaviors that give “socially awkward” robots better “people skills”. Hence, psychological research methods will be crucial in engineering engaging, long-term and motivating interactions between humans and artificial agents (Broadbent, 2017). But how can we solve the problem of designing truly social robots (Duffy & Joue, 2005)? One approach may be to examine a kind of “lowest common social denominator” that helps establish common ground in human-human interaction: namely, interpersonal synchrony. Defined as movements matched in time (Hove & Risen, 2009), interpersonal synchrony has been established as an indicator of social closeness between two individuals, and also a causal factor in enhancing rapport between people (Berniere et al., 1988; Hove & Risen, 2009).

Researchers in human-robot interaction have started taking advantage of the fact that synchrony with another agent may foster rapport (Hove & Risen, 2009). In their proof of concept study, Mörtl, Lorenz and Hirche (2014) equipped a robot with the ability to synchronize its movements to those of human participants during a joint-action pick-and-place task. The authors report that 11 out of 12 participants recognized the adaptability of the robot and 10 participants liked this about the robot. Relatedly, Shen and colleagues (2015) used an information distance algorithm to generate real-time, adaptive motor coordination with the KASPAR2 robot. While the main goal of the experiment was to test the success of the synchrony-promoting algorithm, they also distributed a questionnaire to their 23 participants, inquiring about which of the games (adaptive condition versus non-adaptive baseline condition) they preferred. While most participants preferred the adaptive robot, there was no significant pre- to post- rating difference for their single-item measure of the robots' social capabilities. However, results by Lehmann and colleagues (2015) suggest that movement synchrony of a non-anthropomorphic robot significantly improved participants ratings of the robot's likeability and perceived intelligence.

As Irfan and colleagues (2018) emphasized, when implementing concepts from social psychology to human-robot interaction studies, it is important to establish how reliable and robust these effects are in humans. A recent meta-analysis by Mogan and colleagues (2017) investigated the effect size of interpersonal synchrony on pro-social attitude and behavior. The authors included 42 independent studies that experimentally manipulated synchrony. The researchers found that moving in synchrony had a medium effect on increasing prosocial behaviors (M_{ES} = 0.28), small to medium effects on perceived social bonding and cognition (M_{ES} =0.17) and a small effect on increasing positive emotions (M_{ES} =0.11). However, Mogan et al. (2017) did not take into account a potentially problematic methodological artefact: experimenter bias. In fact, a meta-analysis conducted by Rennung and Göritz (2016) reports that the effect of interpersonal synchrony (here they define synchrony both as "synchronous motor movement and sensory stimulation", p. 169) on prosocial behaviors can be *entirely* explained by a lack of experimenter blinding. They found that the effect of interpersonal synchrony on prosocial attitudes and perceived social bonding was greatly reduced when controlling for experimenter blinding but remained significant.

Similar to the abundance of synchrony manipulations used in the field (L. Cross et al., 2016), no underlying mechanism is generally agreed upon (Mogan et al., 2017). However, Rennung and Göritz (2016) remark that all potential explanations share a common trait: “[interpersonal synchrony] is a rewarding experience” (p. 169). Wheatley and colleagues (2012) hypothesize that moving in sync with another individual may engage the brain’s reward system, which in turn may incentivize further social interactions. This idea is closely related to the theory of social motivation, as proposed by Chevallier and colleagues (2012). These scientists highlight two main components of social reward: liking and seeking of social cues. Empirical support for the theory that interpersonal synchrony may be connected to reward comes from Kokal and colleagues’ (2011) study on synchronized drumming. For participants who acquired the drumming rhythm easily before the scanning session, activity in the caudate nucleus was enhanced during synchronous drumming, which furthermore predicted later prosocial behavior towards the experimenter (who was blind to the manipulation). All in all, a possible underlying social reward mechanism may be what promotes the positive interpersonal effects of synchrony, thus highlighting the need to investigate interpersonal synchrony in conjunction with social motivation.

The goal of the present double-blind study was to investigate whether interpersonal synchrony with a robot improves social motivation towards the robot. We hypothesized that moving in sync with the robot would improve its likeability, analogous to the findings of Lehmann and colleagues (2015), and, based on Chevallier’s Social Motivation Theory (2012), would increase the motivation to interact with the robot, as measured by the number of questions participants chose to ask the robot during a free interaction.

3.2. Methods

Data statement. We report all measures in the study, all manipulations, any data exclusions, and the sample size determination rule. The data and the R analysis script are publicly available via the OSF [link].

Participants. We aimed to recruit the highest number of participants within the testing period (February to April 2018). Initially, the sample consisted of 71 participants. Four participants were excluded from further analysis due to large error rates (losing the metronome more than 30 times, see experimental procedure below) on the task, and four more had to be excluded due to missing data on the Godspeed questionnaires. Two participants were excluded because they reported studying computer science, and one participant was excluded due to reporting a diagnosis of Autism Spectrum Disorder. 11 participants were excluded, as they failed the manipulation check of correctly perceiving synchrony or asynchrony. Four additional participants were removed after completing statistical checks before analyses (see data analysis, below). The final sample consisted of 45 participants. The subjects' ages ranged between 18 and 31, with an average of 20.51 years (*SD*=2.69). Of the 45, 30 were female. Ethical approval was obtained from the Bangor University ethics review board (2018-16221). All subjects provided written informed consent prior to taking part and were reimbursed for their participation either by payment or course credit. Participants were naïve to the goal of the experiment.

Robotic Platform. For the experiment, a Pepper robot was used. Pepper is a 1.2m tall, commercially available humanoid robot from SoftBank Robotics (Tokyo, Japan). Pepper features 20 degrees of freedom and runs a Linux operating system programmable using NAOqi libraries with Python or C++. The robot can run in an automatic animation mode and a controlled animation mode. For the experiment, the controlled mode was used (sometimes referred to as the 'Wizard of Oz' mode). The controlled mode allows full command over movement and speech, where it only acts as instructed by the experiment program, rather than by its inbuilt AI.

Dependent Measures. Participants were asked to assess likeability, anthropomorphism and perceived intelligence of the robot via the three Godspeed subscales of the same name (Bartneck et al., 2009). The items were presented in a scrambled order, as recommended by the authors. All subscales consist of 5 items, which are structured as a 5-point semantic differential scale (for example: “like-dislike”, “machinelike-humanlike”, “unintelligent-intelligent”). The behavioral measure of social motivation was a list of questions provided to the participants, including such questions as “How are you?”, “Are you a boy or a girl?” and “Are you intelligent?” (Appendix A). The number of questions asked was used as a proxy for social motivation.

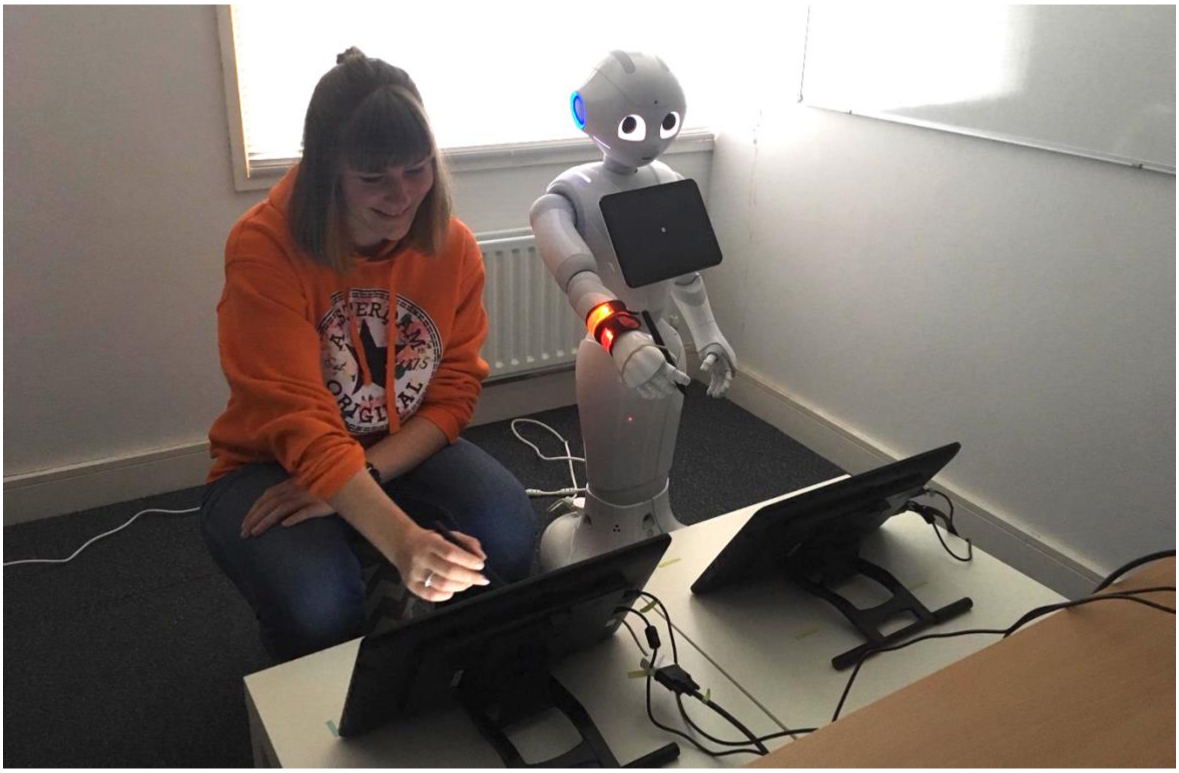


Figure 4 - The set-up for the drawing task.

Experimental Procedure. Upon arrival, participants received information about the experimental task and provided informed consent. Next, they filled out questionnaires relating to their demographic information and trait attitudes towards robots (Nomura, Kanda, & Suzuki, 2006; Syrdal, Dautenhahn, Koay et al., 2009). The Negative Attitude towards Robots Scale (Nomura et al., 2006) for example asks participants to rate statements such as “I would feel uneasy, if robots really had emotions” on a five-point scale from “strongly disagree” (1) to “strongly agree” (5). Then they met Pepper, the robot, who introduced itself as a member of the University research department and invited participants to take a seat next to it. Importantly, the experimenter was blind to which condition the participant was randomly assigned to. The blinding was ensured via a room divider, hence, at no point during the synchrony manipulation could the experimenter see the movements of the robot or the participant.

The two between-subjects experimental conditions involved drawing either in sync or out of sync with Pepper. We modelled our task after Hove and Risen (2009). In their study, participants were following a visual metronome (a rising and dropping bar), which resulted in them tapping either in synchrony or out of synchrony with a confederate (Hove & Risen, 2009). Similarly, we used a visual metronome (a small circle moving along a larger circular trajectory) and instructed participants to follow its movement with a pen. The practical reason for choosing this task was that it gave us a high degree of control of the participants’ movement, without explicitly asking them to synchronize with the robot, a potential confound. In the synchrony condition the metronome was linked to the movement of the robot, whereas in the asynchrony condition the robot was moving approximately 2.5 times as fast along the circle shape as the participant. Participants received the instruction from the experimenter that the goal of the task was to follow the moving target as closely as possible and deviate from it as little as possible. While participants followed the moving target with the drawing pen on the tablet, the robot (due to the technical constraints of it not being able to hold a pen), performed the drawing motion with some distance to the screen (Figure 4). The tablet in front of the robot was always turned off- participants were told that a film on the screen was used to prevent them from getting distracted from their task. When using the drawing pen, participants could see that the pen has indeed a wireless function, but they were always encouraged to keep the pen on the tablet, to minimize the chance of losing the visual metronome. To give a plausible justification for the task, participants were told that the experimenters were looking to investigate the effect of robotic presence on task performance.

After an initial practice round was completed, participants received the additional instruction of monitoring an LED strip on Pepper’s right arm, similar, but not identical, to the one seen in Figure 4. They were told that the LED lights would change colors randomly and they would be probed to report the color changes. However, due to technical difficulties with controlling the LED lights via a remote control, we only report a descriptive graph (Appendix A). Each experimental block consisted of four repetitions around the circle shape, resulting in four circular arm movements per block. After three experimental blocks of the drawing task, the participants filled out the three Godspeed subscales (Bartneck et al., 2009), which were presented to them via the drawing tablet screen. They proceeded with three more experimental drawing blocks.

Finally, they received the instruction via their tablet that the main part of the experiment was over, and they now had the chance to get to know the robot better. They were also informed that this part of the study was optional and that they would not be compensated by research credits or money for the time spent talking to the robot. Then they picked up the piece of paper containing the questions, took a seat opposite to the robot and asked the robot questions, whose answers were Wizard-of-Oz controlled by the experimenter behind the room partition. Then, participants filled out a manipulation check probing them for suspicion and asking about perceived synchrony. Overall, the task took 12 minutes to complete (2 minutes per experimental drawing block) and completing the entire study took roughly 45 minutes.

Data analysis. We conducted a MANOVA on the Godspeed subscales, as this analysis accounts for the relationship between the outcome variables. Before the analysis, multivariate assumption checks were conducted. The Mardia skewness and kurtosis tests confirmed multivariate normality. Via Malanobis distance, four multivariate outliers were identified and removed. Moderate correlation between dependent measures was confirmed after running pairwise correlations. Bartlett's test was not significant, indicating homogeneity of variances. Furthermore, a non-significant Box's M test suggested homogeneity of the covariance matrices. A one-way multivariate analysis of variance (MANOVA) was conducted to investigate the effect of synchrony on the robot's likeability, anthropomorphism and perceived intelligence. Furthermore, Welch's Two Sample t-test was used to examine how the synchrony manipulation affected the participants' social motivation. However, the manipulation check showed that a rather large proportion of the participants in the asynchrony condition had perceived to be in sync with the robot (n=10) and one participant in the synchrony condition had failed to perceive this (*n*=1). Based on this insight, participants who had failed to correctly perceive the manipulation were excluded, resulting in *N*=45 participants (henceforth ‘original group split’). A second group split based on perceived synchrony was performed (henceforth ‘perceived groups’), and within the context of exploration, the above analyses were repeated (*N*=56). This second group split on the basis of participants’ synchrony beliefs was investigated, since previous literature showed that top-down beliefs about a robot’s behavior play an important role in agent perception, over and above bottom-up cues (Klapper et al., 2014; Cross et al., 2016).

3.3. Results

Original group split. The one way MANOVA showed no significant differences between groups on the dependent measures: Pillai’s *V*=.07, *F*(3, 41)=.96, *p*=.42. There was no significant difference between the groups on the measure of social motivation: *t*(41.49)=-.45, *p*=. 67, *d*=-.13. These results are visualized in Figure 5. Synchrony did not lead to increased liking or social motivation towards the robot.

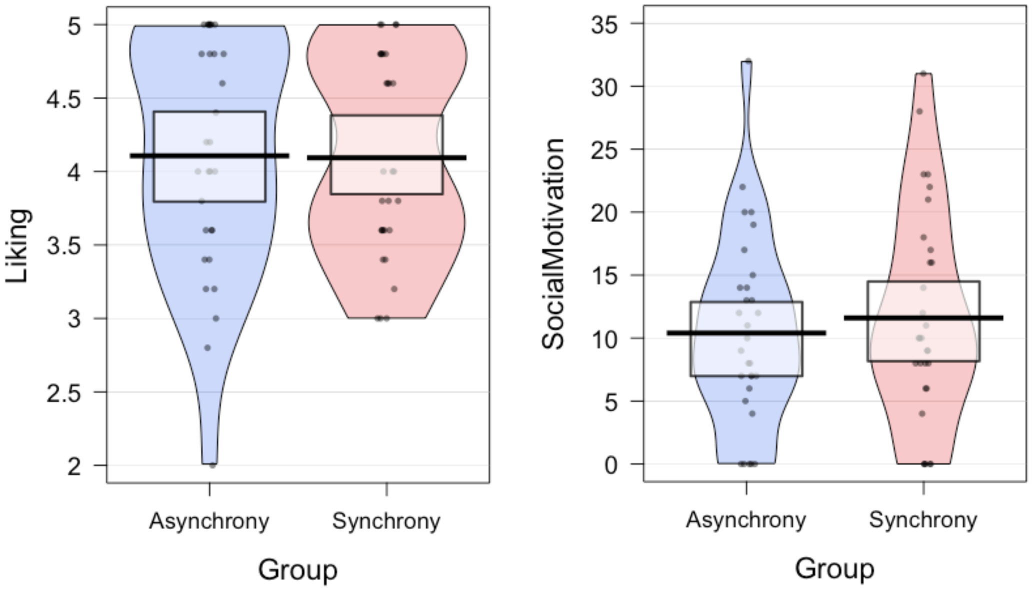


Figure 5 - Experimental groups.

The plot on the left-hand side depicts the groups' ratings on likeability of the robot. The graph on the right depicts the distribution of number of questions participants asked the robot (*N*=45, *n*= 19 in the asynchrony group, *n*=26 in the synchrony group). The plots depict the raw data, the central tendencies and densities, and the 95% highest density intervals.

Perceived groups. The second one way MANOVA showed also no differences, when the groups were split on perceived synchrony: Pillai's *V*=.11, *F*(3, 52)=2.05, *p*=.12. In addition, there was no significant difference between the perceived groups in social motivation towards the robot: *t*(39.24)=-. 26, *p*=.60, *d*=-.15.

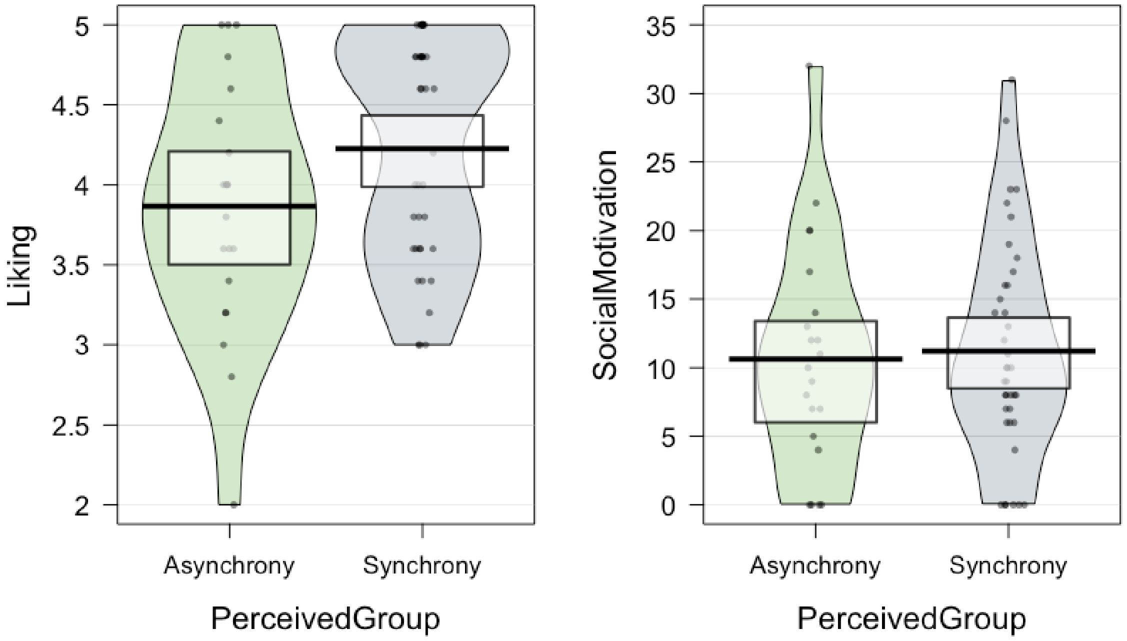


Figure 6 - Perceived groups.

On the left, the likeability ratings are shown for subjectively perceived synchrony with the robot. Individuals, who were in the asynchrony condition, but reported to have been in sync with Pepper were combined with those, who were objectively in sync with the robot. On the right, again the number of questions asked are shown, this time for perceived groups (*N*=56, *n*=20 in the asynchrony group, *n*=36 in the synchrony group). The plots depict the raw data, the central tendencies and densities, and the 95% highest density intervals.

Likeability ratings and social motivation of the perceived groups are depicted in Figure 6. Perceived synchrony did not lead to an improved perception of Pepper or towards an increased motivation to ask the robot questions.

3.4. Discussion

In this study, we investigated the effect of experiencing interpersonal synchrony with a humanoid robot on its likeability and participants' social motivation towards the robot. Contradictory to our hypotheses, participants who moved in sync with Pepper did not rate the robot as more likeable, intelligent or humanlike than participants who performed the task out of sync with it. Participants in the synchrony condition did not show stronger social motivation towards the robot, as indexed by the amount of questions they asked the robot in a voluntary interaction after completion of the main task.

One critical but interesting observation were the differences in experimentally manipulated and subjectively experienced synchrony. One third of the participants who were assigned to the asynchrony group reported that they believed they were moving in sync with Pepper. Given this finding, it may be that the experimental manipulation of synchrony was either too subtle or too short to fully immerse participants in the experience and to produce the hypothesized beneficial effect on rapport between synchronizing agents. Indeed, findings reported by Lehmann and colleagues' (2015) suggest that movement synchrony should positively impact self-reported likeability of a synchronous robot. However, an important difference between the study reported here and their experiment was that in their videos, the robot was making goal-direct movements towards a person. They defined "positive synchrony" as the robot shifting its "gaze" towards the movement of a human agent, who was arranging flowers in a vase. In contrast, in our experiment, Pepper was making goal-directed, synchronous movements reacting to the task, and not the participant. Hence, this was a markedly less social context, than reacting to the movements of the other interaction partner.

In addition to the potential necessity of adaptivity in synchronous interpersonal movement, Lorenz, Weiss and Hirche (2016) argue that in order to reap the benefits of synchrony in social interactions with robots, the human interaction partner needs to attribute a mind to the robot. This idea is consistent with research by (Wiese et al., 2012), which shows that top-down beliefs about an agent's intentional stance can influence basic attentional mechanisms. Even though we assessed trait negative attitudes towards robots, we did not include a self-report or behavioral measure of mind attribution. While Pepper introduced itself before starting the drawing task, it remains unclear how much mind and intention the participants attributed to the robot. In addition to these factors that could have adversely affected the hypothesized positive influence of interpersonal synchrony, we saw a ceiling effect of likeability of the robot – in both groups, Pepper was rated as very likeable.

More questions remain regarding why the synchrony manipulation did not impact participants' social motivation towards Pepper. One possible explanation for this result could be that counting the amount of questions the participants chose to ask the robot may have been too crude a measure to pick up any small to medium sized effect we expected from a synchrony manipulation. Stronger motivational factors, such as the desire to finish an already long experiment, may have interfered with subjects' desire to spend time with the robot. In addition, previous experiences with the robot might have influenced their behavior, with participants lacking any experience perhaps showing stronger curiosity to interact with Pepper or a lack of familiarity affecting the mind perception of the robot (Müller et al., 2011). This lack of sensitivity of the behavioral measure highlights an important gap in readily available, objective, dependent measures in social robotics. Behavioral and neuronal measures offer objectivity, which self-report measures are not able to provide, due to inherent reporting bias and social desirability effects. Drawing on established and validated measures from cognitive (neuro)science might help us to bridge this gap (Wiese et al., 2017). Future research in interpersonal synchrony with robots should invest in the implementation of these behavioral and neuroscientific dependent measures, to complement the limitations of self-report and enable more precise triangulation of the mechanisms and consequences of social affiliation via synchrony. Future experiments should further include a positive control to ensure the synchrony manipulation works as expected in human-human interaction and additional loops of control to ensure that the synchrony manipulation is sufficiently immersive and salient. A final limitation we would like to highlight is the fact that given the rather high number of participants we had to exclude, the sample size may have been too small to show the expected small to medium effect size of a synchrony manipulation on perception of and behavior towards the robot.

Following the tenets of the recent HRI'18 workshop "What Could Go Wrong: Lessons Learned When Doing HRI User Studies with Off-the-Shelf Social Robots?", below we summarize the insights gained as psychologists conducting experiments with commercially available robots, such as Pepper.

The Pepper robot as an experimental confederate: lessons learned.

Our initial motivation was to use the most natural, and most autonomous robotic behavior available. However, we quickly noticed in preceding pilot experiments that even little robotic movements away from the participant (due to it orienting to the experimenter's voice behind the room partition), were interpreted as rejection, and especially the faulty behavior of the robot during the free interaction period (due to volume or accent issues), would obstruct the question asking scenario significantly. As such, we used an experimenter-controlled, Wizard-of-Oz setting with gaze lock implemented, to ensure it would always face the participant during the introduction and free interaction period. Furthermore, we found it useful to use Pepper's "alive and breathing" mode between experimental drawing blocks, as the change from complete stillness to the drawing motions might have been perceived as too uncanny. Further, when employing a humanoid robot in a psychology-informed synchrony experiment, we recommend facilitating a salient experience of synchronizing with the robot, to ensure that experimental results are driven by the manipulation and not the lack of synchrony immersion.

In conclusion, we did not find that orchestrated synchrony, here induced via a drawing task with a physically present embodied robot, improved the rapport between participants and the robot. Future experiments will help to further elucidate the relationship between synchronous behavior and social affiliation toward robots by including both behavioral and neural measures of social motivation.

3.5. Acknowledgements

We would like to thank Bishakha Chaudhury for her invaluable help with programming the drawing task and the robot. Furthermore, we would like to express our gratitude towards Veronica Diveica, Leorah Lasrado, Alex Gillespie and Abdulla Almusalam for their assistance with the data collection. This research was funded by the ERC Starting Grant ‘SOCIAL ROBOTS’ (ERC-StG-2015-62770 to E. S. C.).

Chapter 4 Faces do not attract more attention than non-social distractors in the Stroop task

This chapter is an exact copy of the following manuscript under review:

Chapter 4 - Henschel, A., Bargel, H., & Cross, E. S. (2020). Faces do not attract more attention than non-social distractors in the Stroop task. Accepted pending minor revision at *Collabra: Psychology*. https://psyarxiv.com/pbfny

4. Abstract

As robots begin to receive citizenship, are treated as beloved pets, and given a place at Japanese family tables, it is becoming clear that these machines are taking on increasingly social roles. While human-robot interaction research relies heavily on self-report measures for assessing people's perception of robots, a distinct lack of robust cognitive and behavioural measures to gauge the scope and limits of social motivation towards artificial agents exists. Here we adapted Conty and colleagues' (2010a) social version of the classic Stroop paradigm, in which we showed four kinds of distractor images above incongruent and neutral words: human faces, robot faces, object faces (for example, a cloud with facial features) and flowers (control). We predicted that social stimuli, like human faces, would be extremely salient and draw attention away from the to-be-processed words. A repeated-measures ANOVA indicated that the task worked (the Stroop effect was observed), and a distractor-dependent enhancement of Stroop interference emerged. Planned contrasts indicated that specifically human faces presented above incongruent words significantly slowed participants' reaction times. To investigate this small effect further, we conducted a second experiment (N=51) with a larger stimulus set. While the main effect of the incongruent condition slowing down the reaction time of the participants replicated, we did not observe an interaction effect of the social distractors (human faces) drawing more attention than the other distractor types. We question the suitability of this task as a robust measure for social motivation and discuss our findings in the light of recent conflicting results in the social attentional capture literature.

4.1. Introduction

Glancing upon Giuseppe Arcimboldo's famous 16th century artwork "Air", a collection of colourful birds transforms into the side profile of an elegant man. The effect Arcimboldo cleverly applied to many of his paintings is also known as *pareidolia*, which describes the illusory perception of human faces in random patterns. This tendency is not only capitalized on in the arts, online communication, and product design, but also in research, where variations on the visual illusion are used to investigate mechanisms of face perception (Bubic et al., 2014; Guido et al., 2019; Martinez-Conde et al., 2015; Pavlova et al., 2018; Robertson et al., 2017; Wodehouse et al., 2018).

While the origin of the pareidolia phenomenon is somewhat contentious (with explanations ranging from "visual false alarms" to reflecting a deeply ingrained need for social contact), it points to the fact that human faces have a unique status in our visual environment (DiSalvo & Gemperle, 2003; Wodehouse et al., 2018; Zhou & Meng, 2019). From birth, babies exhibit a preference for gazing at faces compared to scrambled faces, with a bias for gazing at others' eyes developing within the first year of life (Hessels, 2020). Replications of a seminal eye-tracking study by Yarbus (1967) confirm that participants invariably have a gaze preference for people, faces and eyes (DeAngelus & Pelz, 2009). Faces are a rich source of information, giving insight into another person's emotions, their intentions, and their personality traits. Willis and Todorov (2006), for example, have shown that the proverb "you only get one chance to make a first impression" is grounded in empirical truth. They found that participants were able to make reliable trait judgements on attractiveness, likeability, trustworthiness, competence and aggressiveness within split seconds. In yet another study, perceivers were capable of deducing the social class of unfamiliar faces above chance level, highlighting the importance of face perception and its potential societal impact (Bjornsdottir & Rule, 2017).

An integrative theoretical account on the relative importance of social cues, such as faces, by Chevallier and colleagues describes social motivation by means of three main components: social reward, social maintaining, and social orienting (2012). Interactions with others, the authors argue, are inherently rewarding, relationships are driven by our goals to maintain and improve them, and social cues are thus prioritized. The authors propose that social motivation is determined by specialized biological processes, which developed due to an evolutionary advantage of collaborating with other humans. Thus, social information in the form of facial cues is thought to be extremely powerful in terms of claiming attentional resources, increasing our chances for improved coordination and cooperative work with others (Chevallier et al., 2012).

Given their prioritization in our visual environment, it is unsurprising that faces have been the central focus of many visual attention studies. Collectively, these studies point towards faces ranking above objects in capturing automatic attention. Using a change blindness paradigm, Ro, Russel and Lavie (2001) found that participants detected changes in temporarily presented faces more quickly than changes in any other object. This effect disappeared when the face stimuli were inverted. Automatic attentional capture by faces was further investigated by Theeuwes and Van der Stigchel (2006), who critized that Ro and colleagues' (2001) results could have been due to merely a preference for attending to faces, and not reflective of truly exogenous attentional capture. In their inhibition of return paradigm, these authors found evidence for automatic attentional capture induced by faces as compared to object stimuli. The authors observed a delayed gaze response towards locations that had previously shown a face and reasoned that this represented true attentional capture by faces, rather than difficulties with disengaging attention from them. Bindemann and colleagues (2007) sought to understand whether attentional capture by facial cues could be entirely determined by their salience, or whether this effect is also modified endogenously, by participants' own volition. As a matter of fact, participants were able to direct their attention away from faces towards objects when these were more predictive of the cued target location in a dot-probe paradigm. However, the authors claimed an overall face bias persisted, with participants showing greater ease at directing attention to predictive faces versus predictive objects. Experiments by Langton and colleagues (2008) further affirmed the notion that attentional capture by faces is automatic and involuntary. Searching a visual array for a butterfly was slowed by the presence of an "additional singleton", a task-irrelevant face. Here, the authors concluded that humans became consciously aware of faces before any other none-face item. Overall, a large body of evidence suggests that social attentional capture by facial cues is a robust phenomenon, providing evidence for the putative social orienting pillar of the social motivation model.

Beyond seeing faces in oddly shaped clouds, Martian craters or pieces of burnt toast, we also encounter deliberate pareidolic design when we interact with humanoid robots (DiSalvo et al., 2002; DiSalvo & Gemperle, 2003; Wodehouse et al., 2018). Due to the face's role in communicating emotions, and more generally, facilitating social interactions, the design of human-like (or at least human-readable) robot faces has attracted considerable attention and investment in the domain of social robotics. A key driver behind humanoid robot design is the desire to build a believable social agent, while mitigating the potential damaging effects an overly human-like appearance could have on the user (DiSalvo & Gemperle, 2003). Thus, in order to avoid an uncanny experience, or over-promise on the robot's functionality, a popular design choice for socially assistive robots is a humanoid face with simple geometric shapes alluding to familiar, human features (Kalegina et al., 2018). Indeed, when participants were asked to rate the humanness of humanoid robot heads, only a few features accounted for more than 62% of variance: the eyes, eyelids, nose and mouth (DiSalvo et al., 2002). This is in line with a study by Omer and colleagues, which mapped the features that contributed to the global gestalt of pareidolia faces, identifying the eyes and the mouth (2019). Robots' facial cues are viewed as one of the crucial four dimensions in driving human-likeness ratings, and in a survey of humanoid robots, 87.5% had at least some facial features (DiSalvo et al., 2002; Phillips et al., 2018). It is of note that when establishing an impression of animacy, viewing the face as a whole is crucial, with participants being more hesitant to make judgements about the presence of mind in an agent when viewing cropped facial cues in isolation (Looser & Wheatley, 2010). Hence, and as Geiger and Balas (2020) point out, robot faces, which we have presented here as a special case of intentional pareidolia, constitute a border category of face processing, and while some research exists on attentional capture by pareidolic faces, less is known about the social relevance of robot faces. This question however is crucial, as humanoid robots become increasingly commonplace in modern society, taking on care, companionship and support roles. Hence, an important goal is to develop robust behavioural tasks that probe the relevance of robotic, compared to human, social cues.

Research on pareidolic faces and the extent to which they engage social attentional processes has yielded mixed results so far, with some researchers arguing for the crucial role of top-down information driving the face illusion effect (Takahashi & Watanabe, 2013, 2015), and others providing evidence for a bottom-up account of the phenomenon (Liu et al., 2014; Robertson et al., 2017). Takahashi and Watanabe (2013) investigated reflexive attentional shifts induced by pareidolic faces using a gaze cueing paradigm. The authors found a cueing effect of pareidolic faces, however, this effect disappeared when participants were not explicitly instructed that the presented objects could be interpreted as faces. In a follow-up study, Takahashi and Watanabe (2015) found that face awareness, i.e. perceiving an object (here: three dots arranged as a triangle) as a face improved participants performance on a target detection task. This advantage disappeared when subjects were instructed to detect a triangle target shape, rather than a face target. The authors concluded that despite their identical shape, faces receive prioritized further processing due to top-down modulation of face awareness. On the other hand, a study by Ariga and Arihara (2017) did not find that pareidolia faces captured visual attention when presented as task-irrelevant distractors in a letter identification task. However, when human faces were presented as distractors among a rapid serial presentation of letters, accuracy was significantly impaired. There was no difference between pareidolia faces and their defocused control images for any of the various time lag conditions in the letter identification task. While Ariga and Arihara (2017) conclude that attentional capture by facial cues is exclusively reserved for human faces, yet another study shows that pareidolia faces were able to elicit deeper forms of social engagement, surpassing an initial face detection stage and eliciting further specialized processing. In their study, Palmer and Clifford (2020) presented pareidolic stimuli exhibiting directional eye gaze and found that during a subsequent human direct eye gaze task, sensory adaptation had taken place: the illusory faces influenced the perception of the human face stimuli. This finding is at odds with Robertson, Jenkins and Burton’s (2017) conclusion: these authors claim that their participants’ performance on several pareidolia face detection tasks was unrelated to their performance on face identification tasks, suggesting a functional dissociation and no higher-level face processing taking place elicited by illusory faces.

While the evidence on how deeply illusory faces are perceived as social is mixed, they constitute an ideal control for human facial features in social attentional capture tasks. This also raises the question how deliberate pareidolic faces, such as humanoid robots, might engage our visual attention, as these agents are capable of at least some interactions with the physical world. Some preliminary evidence even exists from an electrophysiological study by Geiger and Balas (2020), which suggests that robot faces were more likely to be perceived as objects, rather than faces when presented in an inversion effect paradigm. The authors found that the face sensitive N170 ERP-component was moderately influenced by robot faces, ranking somewhere between objects such as clocks and real or computer-generated human faces.

The neuronal architecture underlying the prioritization of social cues has been shown to include both cortical and subcortical regions, including the amygdala, the ventral striatum, the orbitofrontal cortex and the ventromedial prefrontal cortex. These brain structures, which are reliably engaged during other types of reward processing as well, seem to be sensitive to, or perhaps even signal, the importance of social aspects of our environment (Schilbach et al., 2011). A formal theory in favour of a specialized subcortical fast track was put forward by Senju and Johnson, who coined the “eye contact effect” (2009). The fast-track modulator model claims that eye contact receives prioritized processing via a subcortical route. To test this hypothesis, Conty and colleagues (2010a) conducted experiments on the distracting effect of social cues while participants were engaged in a cognitively demanding task: the classic colour Stroop paradigm (MacLeod & MacDonald, 2000; Stroop, 1935).

Despite the above reviewed variety of paradigms which probe (social) attentional capture, the Stroop task has proven to be a particularly popular vehicle. Named after the psychologist who discovered the effect, hundreds of studies have shown that naming the ink colour of an incongruent colour word (i.e., the word “RED” presented in green) produces slower reaction times than determining the colour of a control word (the letters “XXX” presented in green). This interference effect, which highlights the fact that task-irrelevant information is processed concomitantly and automatically, has inspired a multitude of extensions, including pictorial, spatial, and social versions (MacLeod & MacDonald, 2000). For example, in the facial-emotional Stroop, participants name the ink colour of emotional, compared to neutral faces, which are overlaid with a coloured filter. Past research has shown that sad participants and participants with higher trait anger are slower to name the colour of angry versus neutral faces (Isaac et al., 2012; Van Honk et al., 2000; van Honk et al., 2001). Thus, the Stroop task has been validated as a suitable paradigm to assess the distracting power of task-irrelevant information, such as facial cues.

In Conty and colleagues' study (2010a), the cropped eye-regions of human faces with open or closed eyes - in one of two head orientations - were presented as task-irrelevant distractors on top of the Stroop task. The authors found that the interference effect produced by the competition between the automatic processing of word meaning and ink colour was further enhanced in the *direct gaze* condition, regardless of the head orientation. In a follow-up experiment, Conty et al. (2010a) showed participants visual gratings and grey colour blocks as distractors, which the authors argue excluded the possibility that the effect might have been driven by low-level visual properties of the images – as open eyes have an inherently stronger visual contrast than closed eyes. In a third experiment with a new participant sample, they again found no difference between closed or averted eyes when presented as distractors on the task. Conty and colleagues conclude that the salience of direct eye contact was so strong that it tapped into processing resources needed to perform well on the main task: responding quickly and accurately to the target words (2010a).

A later study from the same lab by Chevallier and colleagues replicated and extended the costly eye contact effect (2013). Importantly, the authors tested the paradigm in two groups of children: typically developing boys and a group of male adolescents with Autism Spectrum Condition (ASC). Again, open and closed eyes were presented as distractors above the neutral and incongruent words, however, this time a non-social control condition was added: flower images. As expected, the authors report the Stroop interference effect, where incongruent words significantly slowed participants' reaction times. The typically developing group showed the hypothesized enhanced interference in the social condition (here open and closed eyes were taken together as the 'social' category), while the ASC group showed the opposite effect. However, when investigating only the open versus closed eyes, stronger interference for open eyes was preserved in adolescents with ASC. The authors interpreted their findings as yet another confirmation for the strong salience of task-irrelevant social distractors but remark that their results are limited by their specific stimulus set and invite future studies to investigate other types of social distractors, such as whole faces.

In the current study, we built on their paradigm by testing the extent to which human, robot or object faces capture attention automatically, by presenting them on top of the classic colour Stroop task. We were interested in extending the Stroop paradigm to test a wider variety of social cues in terms of their motivational value, as well as in evaluating the utility of the social Stroop task with robot faces as a valid behavioural task to probe social perception in HRI research.

Hypotheses. In line with a large body of literature on the Stroop interference effect, we expected that incongruent words would slow reaction times in comparison to the neutral target word condition, leading to the classic interference effect (MacLeod & MacDonald, 2000). Based on the findings by Conty et al. (2010a) and Chevallier et al. (2013), as well as the established literature on social attentional capture, we further predicted that the more socially salient a cue is, the more it would lead to enhanced Stroop interference in this conceptual extension of the paradigm. The most socially salient stimuli used in the present study were human faces, which we predicted would increase reaction times in the incongruent Stroop condition. Less salient distractors were the robotic faces, which in theory allow for a more minimal form of social interaction. Even less socially salient distractors, the object (pareidolic) faces, contained facial cues but no capacity for the object to interact with the world in a social manner. Finally, we expected the control images, which held no social relevance whatsoever, to have no effect on reaction times in the incongruent condition of the Stroop task.

4.2. Experiment 1

4.2.1. Method

Preregistration and data statement. The experiment was pre-registered via www.AsPredicted.org. The document can be found at https://osf.io/ky4b7/. We report all measures in the experiment, all manipulations, any data exclusions and the sample size determination rule (Simmons et al., 2012). Data and the R analysis scripts are available (https://osf.io/xyz4m/). Due to copyright restrictions, the full stimulus set is not openly available, however it can be shared upon request.

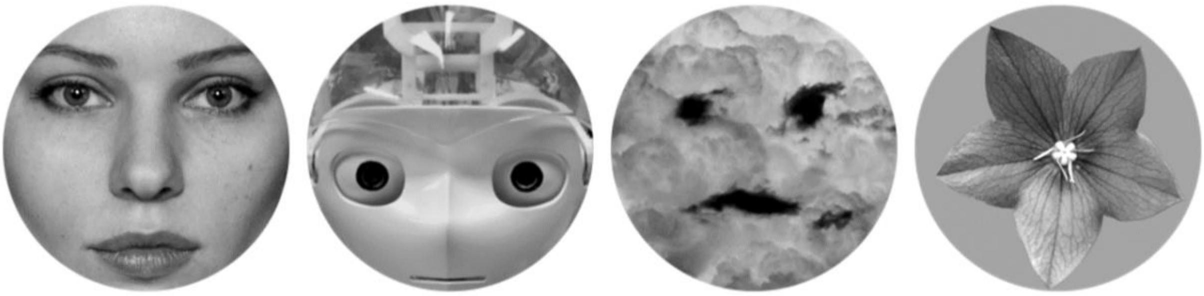


Figure 7 - Stimulus categories.

A representation of the four different stimulus categories: human faces, robot faces, pareidolic faces and the control images, flowers. The human, robot and object distractors all have a direct gaze orientation and show a neutral facial expression. The full stimulus set is available upon request, as individual images are restricted by copyright.

Participants. An a-priori power analysis based on the contrast of interest resulted in a total sample size of 47 participants (*d_z*=0.49, *a*= 0.05, power=0.95, noncentrality parameter = 3.359, critical *t*=1.678, Df=46, actual power=0.95). We recruited 50 participants, however, based on our pre-registered exclusion criteria (diagnosis of ASD and having had a previous interaction with a robot) we excluded 9 participants. Two additional participants had insufficient English language skills, and thus the total number of exclusions was 11. The pre-registered exclusions were made based on participant answers on the experiment questionnaires' self-report items (for example: "Do you have a diagnosis of Autism Spectrum Disorder?" and "Have you interacted with a robot before?"). The other exclusions had to be made in addition, based on the difficulties of the participants with the task. We report a final sample size of

N=39. Of the 39 participants, 26 were female, and reported a mean age of 27.41 years (*SD*= 7.35). Ethical approval was obtained from the University of Glasgow ethics review board (300170224). All participants provided written informed consent prior to taking part and were reimbursed for their participation by payment. As in the original study, the experiment was framed as an experiment on colour perception.

Stimuli. A new stimulus set was built for this adapted version of the Stroop paradigm (Figure 7). The human faces were selected from neutral, frontally oriented facial expressions in the Radboud Faces Dataset and the London Faces Database (Langner, Dotsch, Bijlstra, et al., 2010; DeBruine & Jones, 2017). The robot and object faces, as well as the flowers, were selected from Google, with the aim to include only neutral, frontally-oriented faces. The rationale behind including only neutral faces was that emotional facial cues have been shown to draw attention, especially in comparison to neutral facial expressions (Pessoa et al., 2002; Theeuwes & Van der Stigchel, 2006; Vuilleumier, 2002).

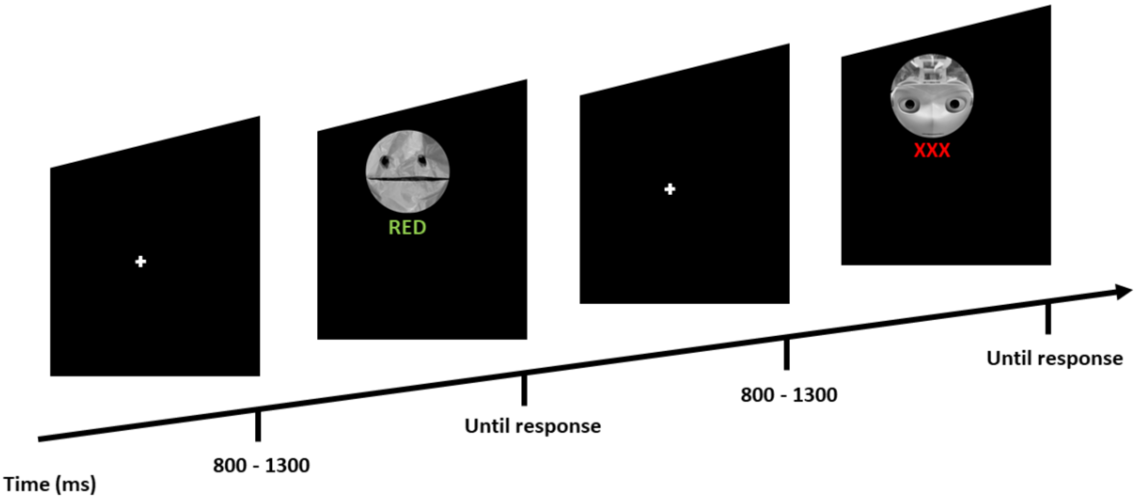


Figure 8 - Schematic representation of a trial time course.

An independent sample rated the first pool of human and robot images, resulting in a pre-selection of more neutrally perceived faces (more details can be found in Appendix C). Twelve unique images were obtained in each of the 4 categories and were edited to achieve a standard round form, mirrored, transformed to grey-scale and averaged according to mean contrast and luminance using the SHINE toolbox in MATLAB (Willenbockel et al., 2010). This resulted in 96 unique images in Experiment 1 (i.e. 24 per each of the four distractor conditions). Since the overall number of trials was 192 (closely modelled on the original study by Conty et al., 2010a), the distractor images were presented twice.

Procedure. Participants were tested in a quiet, dark cubicle on a computer, sitting 50 cm away from the screen. Participants familiarized themselves with the key responses in two training rounds. In the first training, colour-unrelated words (such as “BOWL” or “HAT”) were presented in red, yellow, blue and green ink. Words low in arousal and with a medium valence score from the Affective Norms for English Words (Bradley & Lang, 1999) were selected. In this first practice block, participants received feedback on their performance accuracy and speed, whereas in the second round, the feedback was removed. Each practice block consisted of 48 trials. The experiment was split in 4 blocks, with short breaks after 48 trials. In total, the experiment took 25 minutes to complete.

An experimental trial consisted of a centrally presented fixation cross, whose duration was jittered between 800 and 1300 milliseconds (Figure 8). After the fixation cross, the target word appeared, which extended horizontally over 1° of visual angle, and vertically over 0.5° of visual angle. Directly above the target words, the distractors were presented, extending over ca. 6° of visual angle. The images and word pairs remained on the screen until a response was made. There were equal numbers of incongruent and neutral Stroop trials, and no restrictions regarding the switch between incongruent and neutral trials were put in place, as they were presented randomly. The target word and distractor image pairs were fixed. Due to an error when setting up the PsychoPy experiment (Peirce, 2007), only female human faces were presented in the incongruent condition of the Stroop task, with all the male faces presented in the neutral condition. The object and robot distractor images in Experiment 1 were not one-to-one controlled by their mirror images across the incongruent and neutral conditions.

Statistical analysis (pre-registered). The percentage of accurate responses was calculated and analysed by means of a repeated measures ANOVA. For the analysis of the reaction times, incorrect responses were excluded, as were RTs that were two standard deviations above the mean or below 200ms. As a result, 606 trials (8.09%) were discarded (a detailed breakdown of the trial number per condition can be found in the Appendix C).

We calculated a two-way repeated measures ANOVA with the target (incongruent vs. neutral) and distractor (human, robot, object, flower) as within-subjects conditions. Finally, we conducted planned contrasts. All analyses were conducted in R 3.5.3 (2019), using the {ez}, {psych}, {afex} and {emmeans} packages (Lawrence, 2016; Revelle, 2018; Singmann, Bolker, Westfall, & Aust, 2019; Lenth, 2019).

4.2.2. Results

Accuracy. The repeated measures ANOVA showed a main effect of target, suggesting participants were more accurate in the neutral target word condition: $F(1, 38)= 7.48$, $p=0.009$, $\eta G^2= .03$. However, the overall accuracy was very high (95.72%) and the effect size is considered small, so this was not further investigated.

Table 2 - Mean reaction times and standard errors in milliseconds (Experiment 1).

		Humans	Robots	Objects	Flowers
Incongruent target	*M* (*SE*)	843 ± 11	807 ± 11	815 ± 11	796 ± 11
Neutral target	*M* (*SE*)	753 ± 10	768 ± 11	763 ± 10	760 ± 10

Reaction times. A second repeated measures ANOVA was calculated and as predicted, we saw a main effect of target, with incongruent words slowing down the reaction times of the participants: $F(1, 38)= 39.24$, $p<.001$, $\eta G^2= .03$. This finding confirms that our modified task was still effective at inducing a Stroop interference effect. In addition, we observed a small interaction effect of target x distractor: *F*(3, 114)= 2.69, *p*=.049, *ηG*^{2}= .003. To investigate the difference in reaction times between specific conditions (comparing the effect of the human distractors in the incongruent condition with the flower distractors in the incongruent condition), planned contrasts were computed.

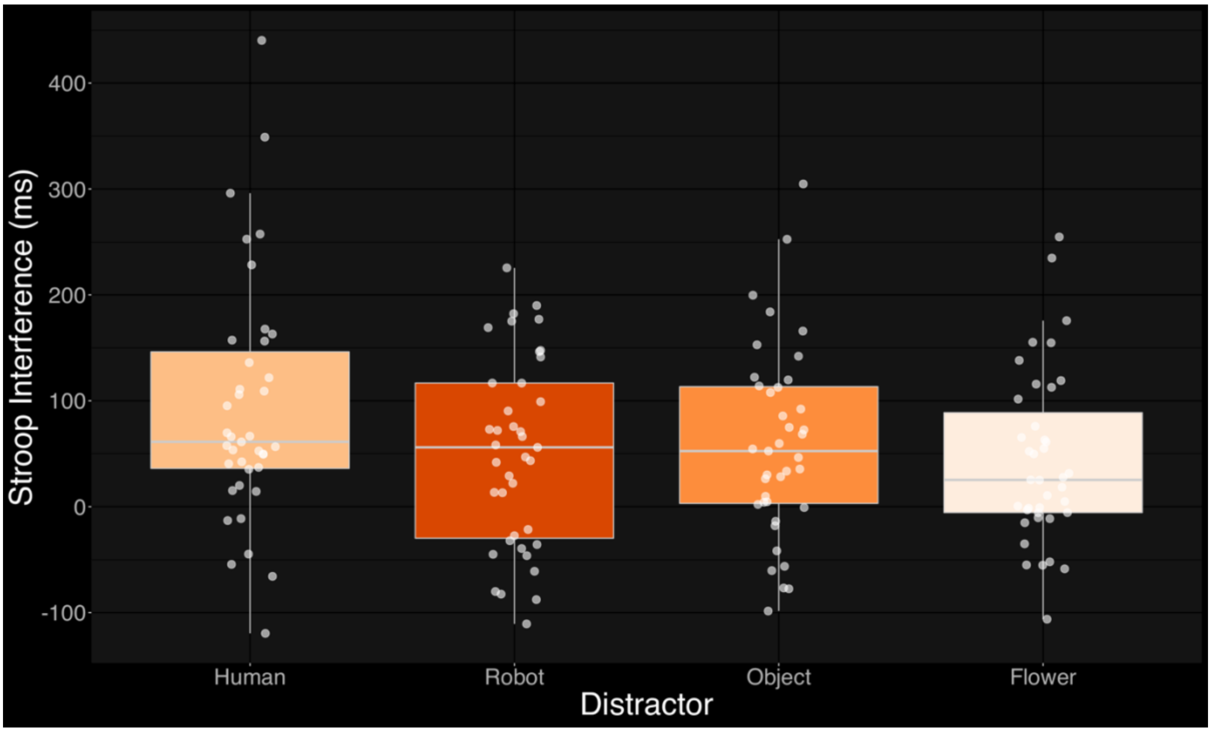


Figure 9 - Results (Experiment 1).

The Stroop interference scores were calculated by subtracting the neutral from the incongruent trials. Here the mean Stroop interference scores are shown for each of the distractor categories in Experiment 1.

They revealed that the human faces were significantly more distracting than the flower images in the incongruent condition: *t*(227)= -2.95, *p*=.004 and drew more attention than the robotic faces as well (*t*(227)=-2.15, *p*=.03), but there was no significant difference to the object faces: *t*(227)=-1.86, *p*=.06. The Stroop interference scores (neutral trials subtracted from incongruent trials) are visualized in Figure 9 and the mean reaction times with standard errors are summarized in Table 2.

4.2.3. Discussion

In Experiment 1 we found an interaction effect in the predicted direction: human faces drew more automatic attention than flower images and robot faces, leading to enhanced interference in the Stroop task. However, the interaction that emerged, as evaluated by the ANOVA, was very small and just above the set significance level (*p*=.049). In addition, due to our conservative participant exclusion criteria, we experienced a larger drop-off in overall subject number than expected. Thus, the experiment was perhaps not adequately powered to detect the effect of interest. Furthermore, we speculated that the effect may have been influenced by the repetition of the distractor images, or due to the described programming error. We next decided to run the same paradigm again, this time recruiting a sufficiently large subject number (accounting for a drop-out rate of approximately 15-20% of participants), presenting both male and female faces in the incongruent Stroop condition, and doubling the number of unique distractors, thus preventing repeated viewing of the stimuli.

4.3. Experiment 2

4.3.1. Method

Preregistration and data statement. We followed the same procedures that were described in our preregistration document, as reported in Experiment 1.

Participants. A new set of participants (*N*=70) was recruited. In addition to the pre-registered exclusion criteria (outlined in Experiment 1 - Method), we added the condition of not having participated in the first experiment. After subject exclusion, 51 participants remained in the sample (39 female). The participants' mean age was 23.24 years (*SD*=6.27). All participants provided written informed consent prior to volunteering for this experiment and were reimbursed by payment. The experiment was approved by the University of Glasgow ethics review board (300180052).

Stimuli. The stimulus set was extended to include 12 new unique images for each distractor condition, which were mirrored and edited in the same way as outlined in Experiment 1. In total, we now had 192 unique distractors.

Procedure. The same experimental procedure was followed as described in Experiment 1. Following the completion of the Stroop task, we also asked participants to rate the unique (unmirrored) distractors based on agency (ability to plan and act) and experience (ability to sense and feel), to establish that the distractor categories were indeed perceived differently, with regard to their varying levels of social saliency. Participants rated each of the 96 images on both characteristics using a sliding scale from 0 to 100 in FormR (Arslan et al., 2019). The questions were derived from Gray, Gray and Wegner's study (2007) on mind perception of different kinds of agents. We used mind perception as a socialness proxy to distinguish between the control condition (flowers), inanimate (robot and pareidolic faces) and agents with a mind (humans). The analysis of the ratings confirmed that the stimulus categories were perceived differently: the human images received the highest agency and experience ratings. A detailed report of the stimulus ratings can be found in Appendix C.

Statistical analysis. We followed the same data cleaning and analysis procedure as in Experiment 1. Incorrect trials were excluded, as well as reaction times below 200ms or 2 standard deviations above the mean (i.e. 1910ms). With this reaction time trimming criterion, we discarded 1061 trials (10.84%). A detailed breakdown of the number of trials remaining per condition can be found in Appendix C.

4.3.2. Results

Accuracy. The repeated measures ANOVA showed no significant main effect of target or distractors, nor any significant interaction effects. Overall, the participants' performance on the task was very accurate again (93.29%).

Table 3 - Mean reaction times and standard errors in milliseconds (Experiment 2).

		Humans	Robots	Objects	Flowers
Incongruent target	*M* (*SE*)	811 ± 10	808 ± 11	809 ± 11	816 ± 10
Neutral target	*M* (*SE*)	723 ± 9	747 ± 9	730 ± 9	735 ± 9

Reaction times. The repeated measures ANOVA on the reaction time data revealed a main effect of target, consistent with the expected Stroop interference in the incongruent condition of the task: *F*(1, 50)=70.31, *p*<.001, ηG^2=.06. Again, this showed that the task worked as expected. The target x distractor interaction was not significant: *F*(3, 150)= 0.36, *p*=.78, ηG^2 =.0003. Planned contrasts were computed using estimated marginal means. No contrast of interest reached significance: there was no difference between human faces and flower images in the incongruent condition: *t*(300)= .094, *p*=.92. The mean reaction times and standard errors are summarized in Table 3 and the Stroop interference scores are visualized in Figure 10.

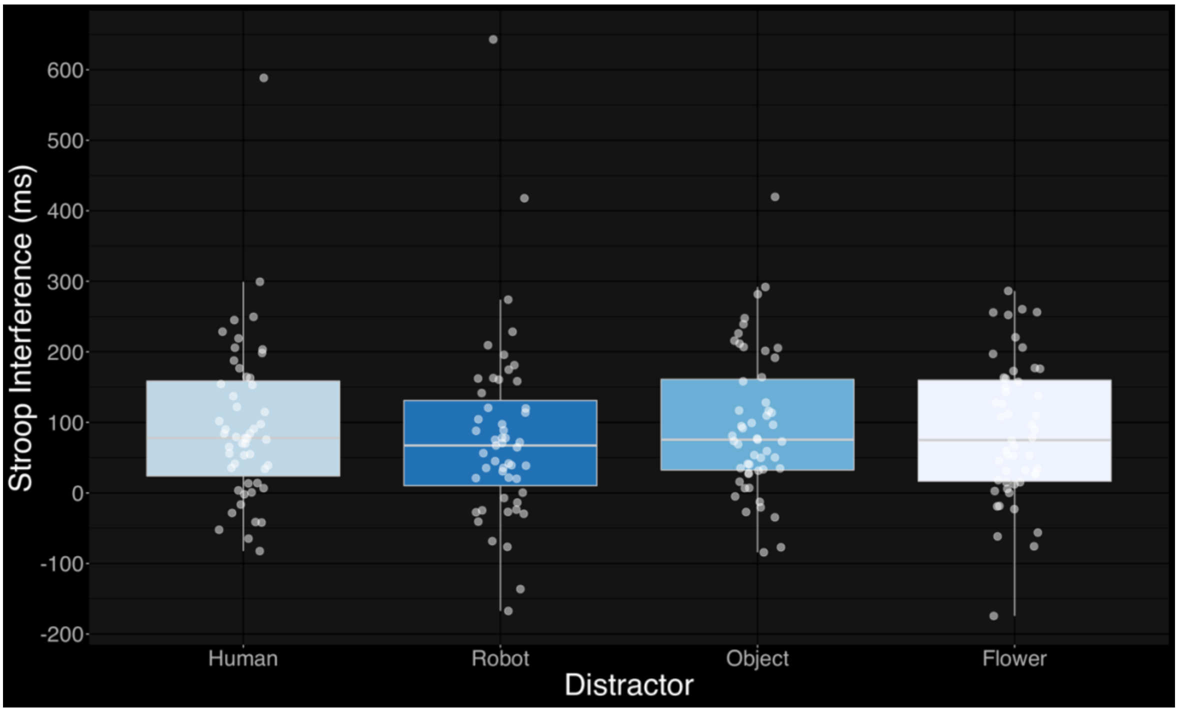


Figure 10 - Results (Experiment 2).

The mean Stroop interference scores (incongruent – neutral conditions) for each of the distractor categories in Experiment 2.

Bayesian regression analysis (exploratory). Given the results of Experiment 2, we explored the extent to which our data provided compelling evidence for the null hypothesis (no enhanced Stroop effect when human faces are presented compared to the control flower condition) by using a Bayesian regression modelling approach {brms} package in R and Stan (Version 2.9.0, Bürkner, 2017), as the null cannot be confirmed with Frequentist statistics.

Following Balota and Yap (2011), we fitted an ex-gaussian distribution to data, as the response shows a strong right-skew (Figure 11). The ex-gaussian distribution is the convolution of the normal and exponential distributions and has been shown to provide a good fit to reaction time data (Balota & Yap, 2011). We included target word and distractor type as fixed effects predictors and included random intercepts and random slopes for each participant in a maximal random effects structure. The same weakly informative prior was applied to all variables, with a Student's *t*-distribution of 3 degrees of freedom, a mean of 0 and a scale of 1. We used the default number of 4 Markov chains, each with 4000 iterations and a warm-up of 1000. This model converged, as supported by R-hat values below 1.01.

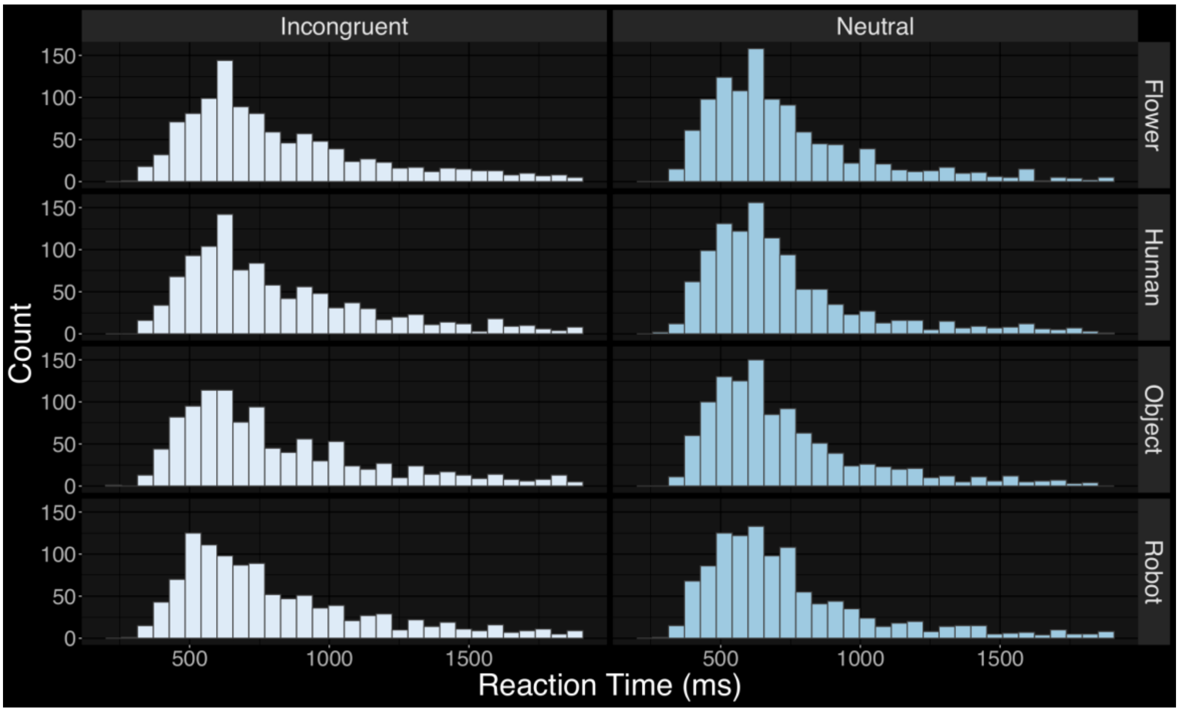


Figure 11 - Reaction time distribution (Experiment 2).
Distribution of the reaction times for each experimental condition (Experiment 2).

We report the estimate (*b*), estimated error (EE) and the 95% credible interval in Table 4 below. The reaction time data was pre-processed in the same way as outlined in the data analysis section of Experiment 1.

Table 4 - Parameter estimates for the population-level effects of the maximal Bayesian model including random intercepts and slopes per participant.

The beta-values of the parameters (b), estimated error (EE) and credible intervals (CI) are shown (Experiment 2).

Predictor	*b* (EE)	95% CI
Intercept	.76 (.01)	[.74, .78]
Incongruent target	.04 (.01)	[.02, .05]
Human distractor	.00 (.01)	[-.02,.01]
Object distractor	.00 (.01)	[-.02, .01]
Robot distractor	.00 (.01)	[-.02, .01]
Incongruent target x human distractor	-.01 (.01)	[-.01, .04]
Incongruent target x object distractor	.00 (.01)	[-.02, .02]
Incongruent target x robot distractor	.00 (.01)	[-.02, .02]

To decide on the acceptance or rejection of a parameter null value we followed the approach outlined by Kruschke and colleagues (2018). Here, a range of plausible values are considered (indicated by the highest density interval (HDI) of the posterior distribution) and how they relate to a region of practical equivalence (ROPE) around null (Kruschke, 2018). The ROPE thus describes effects that are so small that they can be considered meaningless. In determining the ROPE range, we set the limits following the procedure based on half of what we consider a small effect (Kruschke, 2018). A small effect in our first experiment was an average difference of 47ms between the incongruent social and incongruent control distractor, compared to a difference in 34ms in Conty and colleagues' task and 41ms in Chevallier and colleagues' version (2010a, 2013). Choosing the most conservative small effect, we set the ROPE limits to [-.017, .017].

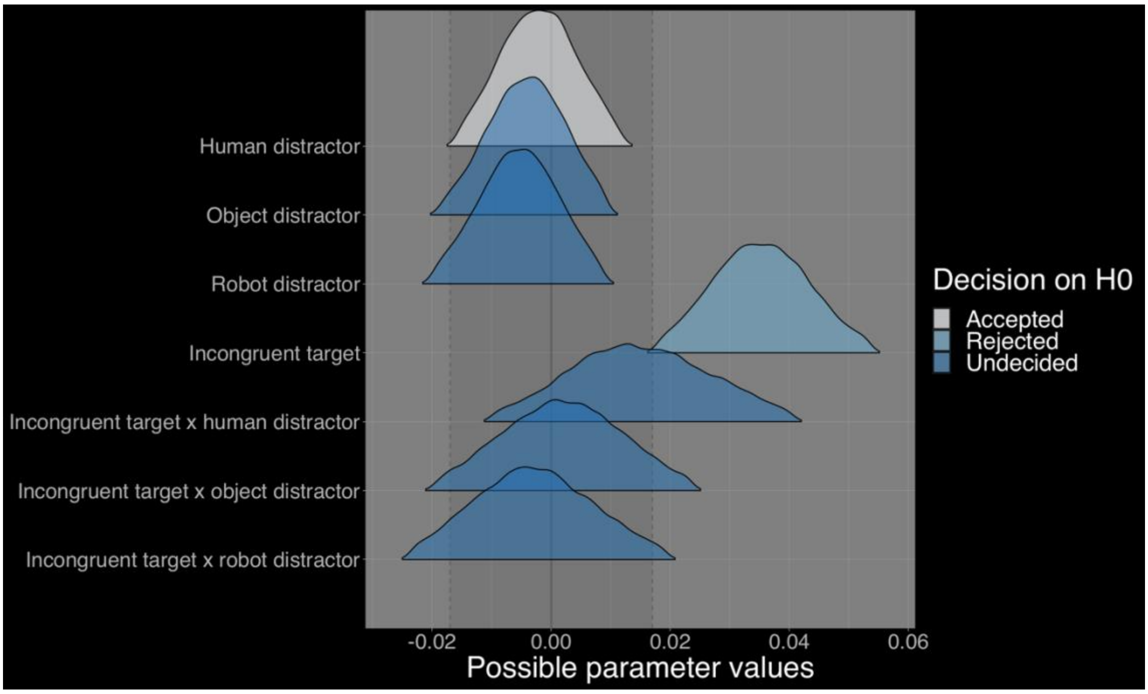


Figure 12 - Region of practical equivalence with zero analysis (Experiment 2).

The region of practical equivalence (with zero) is shaded in grey. The effect of interest (the incongruent target with the human distractor image) is marked in dark blue as undecided (Experiment 2).

As depicted in Figure 12, the ROPE approach here does not offer a straightforward decision on the null hypothesis, even though zero is included in the range of credible parameter values, a small part of the HDI lies outside of the ROPE region for the effect of interest (slower reaction times for human distractors in the incongruent condition).

In summary, in defining our Bayesian regression model, we have increased the uncertainty of our estimates by including more random variance in the form of subject-level random effects. This increased uncertainty is expressed in Figure 12. Based on the ROPE analysis, we cannot definitively support the null hypothesis. However, considering that zero is contained in the 95% interval of credible values of the parameter’s posterior distribution, the evidence for an effect is not very strong, and if real, goes in the opposite direction: -10ms [-10, 40].

4.4. General discussion

Across two experiments, we investigated how distracting faces with varying degrees of social salience were during a classic Stroop paradigm. Contrary to predictions derived from the fast track modulator model by Senju and Johnson (2009), and previous studies demonstrating robust attentional capture by task-irrelevant faces, we did not consistently observe the most salient social cues (human faces) leading to greater interference on the Stroop task. While we report a marginally significant interaction in Experiment 1, suggesting stronger distractibility of human faces in the incongruent condition, we caution interpretation of this finding, as we conducted our analysis on a smaller participant sample than planned. Thus, we reran our experiment with sufficient power, where we also used a larger number of unique distractor images. While we again observed the predicted general Stroop effect, the target by distractor interaction disappeared. Bayesian reanalysis of the data does not exclude the possibility of the human distractors influencing reaction times more than the neutral control distractors in the incongruent condition. However, this small predicted effect is likely not very strong. Overall, our findings contradict those reported by Conty and colleagues (2010a) and Chevallier and colleagues (2013), who both found task-irrelevant social cues automatically captured attention. While their findings provided empirical evidence for the fast-track modulator model, which posits that social cues should exogenously and automatically engage attention, we don’t see convincing evidence for this from our study. Our results not only appear counter-intuitive given the previous studies this work was based on, but also within the wider context of the literature documenting the reward value of social cues (Chevallier et al., 2012; Williams et al., 2019; Williams & Cross, 2018).

However, empirical evidence for social distractors always capturing attention is less convincing than the two studies by Conty and colleagues (2010a) and Chevallier and colleagues (2013) suggest. A conceptual extension of their task from the lab of Hietanen, Myllyneva, Helminen and Lyyra (2016) failed to replicate the enhanced Stroop effect by direct gaze in a real-life version of the task. In their study, a confederate was looking at participants directly above a screen, which displayed a colour-matching version of the Stroop task. Hietanen and colleagues (2016) found a main effect of direct gaze speeding up the RTs of the participants, as compared to averted gaze. The authors reconcile their contradictory findings by relating them to the higher arousal produced by their stimuli: eye contact with a real person should be more engaging than pictorial representations thereof. In so-called low arousal contexts, they argue, salient social cues *should* recruit attentional resources and interfere with participants' performance on cognitive tasks. In our experiments, even in a context that Hietanen and colleagues (2016) describe as "low arousal", it is most probable that any social salience effect is practically equivalent to zero.

How can our results then be explained? Of course, the stimuli we presented were more complex than those used in the original studies, so it is possible that the eye-contact effect only holds in (more) simplified contexts. The eye region in our stimulus set appeared smaller than in the original experiments, due to it taking up a smaller percentage of pixels in our distractor images. While the eye region itself was smaller, all our social stimuli (the human, robot and object faces) depicted direct gaze and a frontally oriented face. They only varied in their potential as a social interaction partner. So, if the eye-contact effect were to hold, we should have seen a consistent difference between our most salient social stimuli with direct eye gaze (the human faces) and the neutral control condition (flowers). The fact that our data did not support this pattern is especially surprising given that past studies examining direct gaze have also used full-face stimuli in similar, cognitively demanding tasks (Burton et al., 2009; Conty, Russo, et al., 2010b).

A close look at the social attentional capture literature reveals a variety of methodological issues and contradicting findings across studies investigating faces and facial features as task-irrelevant distractors. Many studies report effects based on very small samples (some as small as 8 participants per experiment; (Ariga & Arihara, 2017; Miyazaki et al., 2012; Sato & Kawahara, 2015), make bold statements based on modest statistical evidence ("the three-way interaction approached significance, *F*(2,76) = 2.46, *p*<.10", p. 1103, Hietanen et al., 2016) or use small sets of distractor images which are repeated across many experimental trials (Bindemann et al., 2007; Theeuwes & Van der Stigchel, 2006). Indeed, some of these problematic confounds have been highlighted and tested by Pereira and colleagues (2019; 2020).

Pereira and colleagues (2019, 2020) systematically controlled for each known confound in the social attentional literature, including the perceived attractiveness of stimuli, low-level features and a list of other stimulus properties. In their studies, the authors utilized the dot-probe paradigm, with faces, houses and scrambled distractor images as task-irrelevant cues. The targets appeared with an equal likelihood at six different locations. Pereira and colleagues found across multiple experiments that faces did not reliably draw attention to their cued location, as indexed by participants' reaction time. In a follow-up Bayesian analysis on one of their experiments, the authors found evidence for the null hypothesis of no reaction time differences emerging for targets appearing at locations that were cued by faces or houses (Pereira, 2019). While a different task was used in these studies, the authors' findings closely align with ours: faces are not reliably capturing attention and impairing the performance on an unrelated cognitive task. Interestingly, in a direct replication of Bindemann and colleagues (2007), using less well-controlled stimuli, the authors were able to replicate the effect of attentional capture by task-irrelevant faces, providing convincing evidence for systematic confounds obscuring the true picture in the existing literature.

More evidence for the variable nature of findings on automatic attentional biasing by social cues comes from a series of experiments by Framorando and colleagues (2016), who, similar to Hietanen and colleagues (2016), also failed to replicate attentional capture by direct gaze, when faces were presented in a stare-in-crowd task paradigm. Based on previous literature on this effect, one should expect that faces with direct gaze should be more distracting than faces with averted gaze. The authors found that straight gaze had a faciliatory effect when it was part of the target of the task, not a task-irrelevant distractor cue. These findings were later extended by the same authors, emphasizing again the task-dependent nature of directly gazing faces, which in this study hinged on the social or non-social nature of the task (2018). These empirical findings echo an fMRI study by Pessoa and colleagues (2002), who investigated attentional capture by *emotional* facial cues. Here, like the fast-track modulator model, a popular theory suggests that a subcortical route gives preference to the processing of emotional facial cues. However, the authors found that brain regions implicated in emotion perception were only active when participants were able to attend to the emotional facial cues, and these same brain regions were not differentially modulated when participants were engaged in a cognitively demanding task. This, the authors conclude, means that attentional resources are in fact necessary to allow the neural processing of emotional facial cues.

While we can reconcile our results with these studies, one may still wonder why social cues, which are thought to be inherently rewarding, failed to engage participants in our experiments in the expected manner (Anderson, 2016). Speaking to this, recent findings on reward-related distractors impairing participants' performance have also called this intuitive hypothesis into question (Rusz et al., 2019). A new meta-analysis suggests that the effect size of studies on reward-related distraction is small, and that findings across reviewed studies are highly variable, with reverse results not being uncommon (Rusz et al., 2020). This dovetails with the contradictory results we have found in the literature of social attentional biasing and which have also been addressed by Pereira and colleagues (2020).

Of course, based on this small number of empirical studies, we do not wish to claim that salient social cues, such as faces, never capture automatic attention in any context. Indeed, there is mounting evidence that overt attention (i.e. eye saccades towards social cues), as opposed to covert attention, which is measured by manual reaction time, is consistently directed towards the eye region of faces (DeAngelus & Pelz, 2009; Hayward et al., 2017; E. J. Pereira et al., 2020). Still, we do wish to challenge the putative fast track modulator model and speculate that when faces are presented as task-irrelevant distractors, they may not be salient enough to draw attention and cognitive processing resources away from the task at hand. Furthermore, we question the suitability of the task as a "proxy for social motivation", as suggested by Chevallier and colleagues (2013, p. 1649).

However, our findings should also be interpreted with the following limitations in mind: over the course of two experiments, we recruited from an ethnically diverse participant pool at the University of Glasgow, while presenting rather homogenous looking human faces, consisting exclusively of Caucasian individuals. Given that the studies we based our experiments on did not explicitly mention or measure this factor, we did not assess ethnic background in the short demographic survey preceding both studies. As such, we cannot test whether this aspect played a role in the missing enhanced Stroop interference effect for the human distractor images.

A further stimulus-based limitation was that in Experiment 1, distractors were not controlled by their mirror and presented twice. Thus, the repeat presentation could have led to a particularly memorable stimulus set. In Experiment 2, the unique distractors in the incongruent condition were controlled by their mirror images. Of course, on the other hand, the repeat presentation of distractor images is common practice in the social attentional capture literature (for example, a set of four unique human and pareidolic face images used for an experiment consisting of 450 trials, (Ariga & Arihara, 2017). Takahashi and colleagues (2013) used stimuli with three unique identities over many trials, and only four unique stimuli in another study (Takahashi & Watanabe, 2015). Theeuwes et al. (2006) presented 12 unique distractor images across 96 trials. To put it differently, based on the conventions of the social attentional biasing literature, it is unlikely that we did not observe the expected effect due to the number of unique distractor images we presented.

Despite our best efforts to only include neutral faces, the emotional content of the social stimuli could not be controlled to a fine-grained degree, as it was limited by the design and availability of the robots and objects that were identified through our Google search. In the emotion rating experiment, which we undertook prior to Experiment 1, the robot faces were not rated as unambiguously neutral as the human faces, even after excluding the outliers. Human faces were selected from the neutral category of the Radboud and London faces database, so these stimuli would have contained inherently less variance in perceived emotionality than the robot and object faces. However, given the scarcity of frontally oriented and high-quality robot and object faces, we chose to operate within those constraints. Moreover, in comparison with other studies on social attentional biasing we were able to control for the following confounds (as outlined in Pereira et al., 2020): size and shape of the distractors, luminance and contrast, distance from fixation, the internal configuration of facial features of the human, robot and object images (i.e. a comparable set of features including eyes, a nose and a mouth in most of the images), as well as the task context.

While this set of experiments constitutes a conceptual extension to face stimuli, rather than a direct replication of the eye contact effect, we kept most other aspects of the experimental procedure identical to the studies we modelled our task on. Based on these studies and the facial attentional capture literature, we would have expected that human faces would be most salient, regardless of the small modifications we made. Indeed, keeping in mind recent calls for more generalisation efforts in psychological science (Yarkoni, 2016), we feel that a conceptual replication adds crucial insight to the field of motivated cognition. Further to the arguments we presented, our question and approach directly relate to the conceptualized fast-track modulator model: we tested and failed to support Chevallier and colleagues' (2013) hypothesis that this effect should generalize to other social cues – like faces - as well.

For future research, our findings have important implications: many researchers in human-robot interaction (HRI) lament the absence of robust behavioural tasks to assess social interactions with robots, especially regarding changes in social motivation towards them (Baxter et al., 2016; Eyssel & Kuchenbrandt, 2012; Henschel & Cross, 2020a). A few research groups have successfully adapted cognitive tasks for HRI, for example the inversion effect (to examine anthropomorphism), and the Posner gaze-cueing paradigm (Wykowska et al., 2014; Zlotowski & Bartneck, 2013). Yet, behavioural tasks that reliably assess social motivation towards robots are still scarce. Based on our findings, a suitable point of departure for future generations of social robotics researchers could be to examine overt attention in preferential looking paradigms or saccadic choice tasks, utilizing eye tracking technology (Crouzet & Thorpe, 2010; Fletcher-Watson et al., 2008), as these effects appear robust (Hayward et al., 2017). Another option could be to implement more natural social interaction tasks and measure attentional engagement and shifts in a similar manner as Hayward and colleagues in their conversational paradigm, in which participants' eye gaze behaviour was recorded with spyglasses and cameras (2017). Interestingly, the authors found that the social attention of participants in a natural context was unrelated to their behaviour in the classic Posner gaze cueing task. Their findings also speak to recent calls in the HRI literature to implement more natural, embodied experiments with robots to test changes in attitudes, behaviours and neural correlates in a more ecologically valid context (Henschel, Hortensius, et al., 2020b).

On a more fundamental level, one should reflect on the issue of small effect sizes to be expected in experimental psychology (Funder & Ozer, 2019; Ramsey, 2020; Schäfer & Schwarz, 2019). Based on the insights of recent large scale replication projects, we can be fairly certain that many established effects in the literature are much smaller than initially presented, if replicable at all (Camerer et al., 2018). One should then question what the smallest effect size is that one would consider interesting. Going forward, researchers should aim to conduct well-powered direct replications and consider expected effect sizes before adapting social motivation paradigms for HRI.

When Arcimboldo originally painted his whimsical portraits in the late 16th century, little did he know that machines today would be endowed with facial features to evoke illusory socialness – a simple, yet effective trick, corroborated by data that show that mechanical and screen-based robot faces are rated as humanlike, friendly, intelligent or in some cases, as uncanny (Chesher & Andreallo, 2020; Kalegina et al., 2018; Phillips et al., 2018; Vallverdú & Trovato, 2016). As our surroundings become increasingly populated by a variety of artificial agents (including robots and virtual agents), an important aim will be to probe how different types of faces are processed, and what we might learn about humans’ intrinsic social motivation toward artificial agents’ faces (Geiger & Balas, 2020).

4.5. Acknowledgements

We would like to thank Lauren Colbert for her assistance with the data collection for Experiment 2 and Prof. Guillaume Thierry and Dr. Kami Koldewyn, who helped improve the design of these studies with their valuable feedback. We also want to express our gratitude towards Dr. Andrew Milne and Eline Smit for sharing their insights into Bayesian regression analysis and to Dr. Andrew Milne and Te-Yi Hsieh for their valuable comments on the final draft of the manuscript.

Chapter 5 Social Cognition in the Age of Human-Robot Interaction

This chapter is an exact copy of the author accepted manuscript of:

Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social Cognition in the Age of Human-Robot Interaction. *Trends in Neurosciences*. https://doi.org/10.1016/j.tins.2020.03.013

5. Abstract

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

5.1. Human Neuroscience as the Icebreaker in a Social Robotics Winter

Human-robot interaction (see **Glossary**) is a young field currently in a phase of unrest. Since the development of KISMET in the MIT Media Lab in the late 1990s, one of the first social robots, significant progress has been made towards engineering robots capable of engaging humans on a social level. Robots that respond to and trigger human emotions not only enable closer human-machine collaboration but can also spur human users to develop long-term social bonds with these agents. While progress in developing increasingly innovative and socially capable robots has advanced considerably over the past decade or so, some have suggested that the field is approaching a **social robotics winter**. Referencing the period of disillusionment following escalating hype surrounding artificial intelligence (S. Natale & Ballatore, 2020), the still-limited social repertoire of even the most advanced embodied robots calls into question the proclaimed “rise of the social robots” (Campa, 2016; Tulli et al., 2019).

With robots failing to deliver on expectations, social interaction has been named one of the ten grand challenges the field of robotics is now facing (Yang et al., 2018). To facilitate progress toward this endeavour, the rich literature of cognitive neuroscience offers many insights into human social behaviour, not only on a surface level, but also relating to underlying functional and biological mechanisms (Chaminade & Cheng, 2009; Hortensius & Cross, 2018a; Wykowska et al., 2016). Both human-robot interaction researchers and neuroscientists working with robots converge in their interest in facilitating smooth and successful social encounters between robots and humans. This joint effort should ultimately enable society at large to take advantage of the often-heralded potential of robots to provide economical care, company and coaching.

In this Opinion, we argue that studying the human brain when we perceive and interact with robots will provide insights for a clearer and deeper understanding of the human side of human-robot interaction, and will thus set the stage for a **social robotics** spring. Our focus on the human side of these interactions, including consideration of the constraints of social cognition, serves to highlight what recent advances in human neuroscience, in terms of method and theory, can contribute to fluent human-robot encounters. The focus of the majority of past studies has been the passive perception of other agents. While this work provides a first step towards characterising social interactions, a focus on perception alone neglects the rich, complex, and dynamic nature of behaviours that unfold during social exchanges in the real world. How can social neuroscience further our understanding of not only perception but also of dynamic relationships with robots? These insights will explain how people view and treat these new agents in relation to humans, pets and other animals, and tools and objects. Moreover, answers to these questions will help us to understand and support resulting societal changes in the domain of care, education, ethics and law. In reflecting on the neurocognitive machinery that supports human-robot interactions, we suggest that focusing on representations of social cognition and how these change during actual and sustained interactions with physically present robots will be important. Moreover, we argue that minimally invasive mobile neuroimaging techniques offer exceptional promise for deepening our understanding of the human side of human-robot interaction. These methods will accelerate human-robot interaction research by incorporating social dimensions into our exchanges with these machines, thus generating crucial insights helpful in meeting the grand challenge of creating truly social robots. After all, roboticists, neuroscientists and robots will all benefit from an improved understanding of human social cognition in an age of robots (Chaminade & Cheng, 2009; Hortensius & Cross, 2018a; Wiese et al., 2017).

5.2. The Origins of Imaging the Human Brain During Interactions with Robots

Human fascination with creating a mechanical self dates back to antiquity, with writers in ancient Greece and ancient China conjuring human-like automata to serve as workers and servants (Broadbent, 2017). In the past century, the type of automaton that has most captured the human imagination (and research and development investment) is robots, with some contemporary models edging closer to the fictionalised ideals that first appeared centuries ago. Concurrent with advances in robotics technology has been the advent and rapid development of human brain imaging technology. This technology has been vital in developing our understanding of the neurocognitive mechanisms that support social behaviour among humans. More recently, the fields of human-robot interaction and neuroscience have begun to intersect, providing new vistas on social cognition during interactions with social robots, with seminal studies investigating motor resonance, action observation, joint attention, and empathy felt towards robots. These studies showcase the diversity of brain imaging modalities involved and the technical advances evident from early human-robot interaction research and provide a starting point for neurocognitive perspectives on these interactions.

One initial study in this domain (Gazzola et al., 2007) probed the flexibility of the **Action Observation Network** and reported that the parts of the parietal, premotor, and middle temporal cortices ascribed to this network respond both to watching humans grasp and manipulate objects as well as an industrial robot arm perform these same actions. These findings were corroborated by an electroencephalography (EEG) study showing mu-suppression over sensorimotor or Action Observation network regions for both robotic and human agents (Oberman et al., 2007). Insights into motor resonance for robotic actions were further replicated and extended when researchers (Cross et al., 2012) reported a series of two fMRI experiments that found the Action Observation Network to be, in fact, more strongly engaged during observation of (unfamiliar) robot-like motion, regardless of whether a human or robotic agent performed the movement. These and other surprising findings (reviewed in Press, 2011) were attributed to greater modulation of the Action Observation Network following greater **prediction errors** due to the unfamiliarity of robotic motion.

While observing robotic movements engages action-related brain areas, questions remain regarding the extent to which human observers also ascribe emotions and intentions to lifeless machines. Past brain imaging studies reveal that humans do indeed show engagement of the **Person Perception Network** when observing emotional expressions of robots (Hortensius et al., 2018b) and interactions between robots and other humans (Wang & Quadflieg, 2014). The circumstances under which similar brain responses linked to empathy might emerge when observing humans and robots in simulated pain (Rosenthal-Von Der Pütten et al., 2014; Suzuki et al., 2015), or when attempting to decipher the intentions of robots (Hortensius & Cross, 2018a), remain an active field of inquiry. An fMRI experiment using the **gaze cueing paradigm** showed behavioural and brain responses linked to **mentalising**, such as enhanced activation of bilateral anterior temporo-parietal junction, only when people believed that another person controlled the robot (Özdem et al., 2016).

5.3. State-of-the-Art of Human Neuroscience Approaches to Human-Robot Interaction

Major strides have been made in applying advances in human neuroimaging technology to studying human-robot interaction in contexts that approximate more naturalistic social interactions. These studies further illuminate not only the flexibility and limits of human social neurocognition when perceiving and interacting with robots, but also some of the challenges and opportunities that roboticists face (and will continue to face) as they develop increasingly social robots. Work in this domain highlights the importance of not only stimulus cues to socialness (i.e., does the agent *look* and *move* like a human or a machine?), but also, and arguably even more importantly, how perceivers' prior beliefs or expectations shape brain responses and behaviour (Cross et al., 2016; Gowen, 2016; Klapper et al., 2014).

Neuroscientists are now also taking advantage of increasingly sophisticated and multivariate analytical approaches to more sensitively probe how the human brain represents robots compared to people (**Box I**). Recent work has applied representational similarity analyses to fMRI data collected when participants viewed three agents (a human, an android, and a mechanical-looking robot) performing different actions (Urgen et al., 2019). Results revealed that different nodes of the Action Observation Network represent distinct aspects of these actions, and these representations appear to be hierarchically arranged. Specifically, occipitotemporal regions coded for lower level action features (such as form and motion integration), while parietal regions coded more abstract and semantic representational content, such as the action category and intention. These findings corroborate related work that examined effective connectivity between these two nodes when participants viewed actions of varying familiarity (Gardner et al., 2015).

Additional recent work highlights important aspects of how the human brain computes and evaluates anthropomorphism (Rosenthal-von der Pütten et al., 2019; Waytz et al., 2019; Wiese, 2018). One study has attempted to evaluate the **uncanny valley hypothesis** using an elegant combination of modelling behavioural ratings and functional connectivity brain data (Rosenthal-von der Pütten et al., 2019). The authors reported a response profile within the ventromedial prefrontal cortex that closely reflected the hypothesised, nonlinear, uncanny valley shape when viewing images of robots and humans rated more or less unsettling. Further modelling demonstrated that a distinct signal originating in the amygdala predicted when participants would reject artificial agents. This finding ties in with another recent study (Waytz et al., 2019) that examined anthropomorphising behaviour among a small group of individuals with rare basolateral amygdala lesions. These individuals were able to anthropomorphise animate and living entities similarly to neurologically intact individuals, but anthropomorphised inanimate stimuli (such as a robot) less than controls. The authors suggest that the limbic system plays a key role in processing signals originating from artificial agents in a social versus non-social manner.

However, mere observation of robots in one-off laboratory studies can tell us only so much about human-robot interaction. Two recent fMRI studies highlight further innovations in bringing together neuroscience, robots, and real-world interactions to advance the fields of social cognition and social robotics collectively. The first paves the way for future social neuroscience studies that incorporate unrestricted social interactions with autonomous agents while simultaneously measuring brain responses (Rauchbauer et al., 2019). The authors describe a framework that allows participants to interact with a conversational agent (a Furhat robot) or a human partner while a multimodal dataset is collected including behaviour (e.g., speech, eye gaze) and physiology (e.g., respiration, neural activity). Initial results show less engagement of specific brain regions playing a role in everyday social cognition, such as the temporo-parietal junction and medial prefrontal cortex, during live human-robot interaction compared to human-human interaction (Rauchbauer et al., 2019). Another study examined the extent to which a prolonged period of time spent socialising with Cozmo, a palm-sized, playful robot, shapes empathic responses to seeing that same robot “in pain” (Cross, Riddoch, et al., 2019). These authors employed pre- and post-socialisation intervention fMRI sessions and measured **repetition suppression** within the **pain matrix** to determine whether a week of daily interactions with Cozmo would shift participants’ empathy toward the robot to look more like empathy for another person, based on neural activity as well as behavioural responses. While this study did not find compelling evidence that a week of socialising with a robot discernibly shifted empathic responses to look more human-like (Cross, Riddoch, et al., 2019), this work nonetheless sets the stage for studying the impact of longer-term interactions with robots on social neurocognitive processes. This area of work is crucial if robots will indeed be taking on more social roles in close proximity to humans in our daily lives and should inform robotics developers on ways to maximise social engagement not just for an hour or during an initial encounter, but over the long term.

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

5.4. How Should we Probe the Neurocognitive Reality of Human-Robot Interaction?

Neural responses, as measured with fMRI and EEG, when perceiving or interacting with robots differ vastly across different brain networks. Generally, activity within the Person Perception Network is not reduced when people observe social robots and other artificial agents compared to people, while activity within the **Theory-of-Mind network** is reduced (Cross et al., 2019; Hortensius & Cross, 2018a). Going beyond differences in neural activation magnitude, future research in this area will be propelled by mapping the neural representation of social cognition when we engage with robots and characterising how these representations change over time (**Box I**).

Many studies examining how humans perceive and interact with robots have focussed on the Theory-of-Mind network and the Person Perception Network. These two networks underlie everyday social cognition and are a suitable starting point to investigate the engagement of social cognitive brain regions when encountering robots. Yet, emerging evidence suggests that other brain regions, including the inferior parietal lobule, play a key role when we engage with social robots (Figure 13). Increased activity in object-selective brain regions has consistently been reported across studies using different robotic agents (Cross et al., 2012; Cross, Riddoch, et al., 2019b; Rauchbauer et al., 2019). It is therefore critical to capture changes beyond the standard Person Perception and Theory-of-Mind networks to provide an unbiased account of human-robot interaction, while simultaneously acknowledging the possibility that the robots are perceived as objects after all, at least in some respect or in certain circumstances.

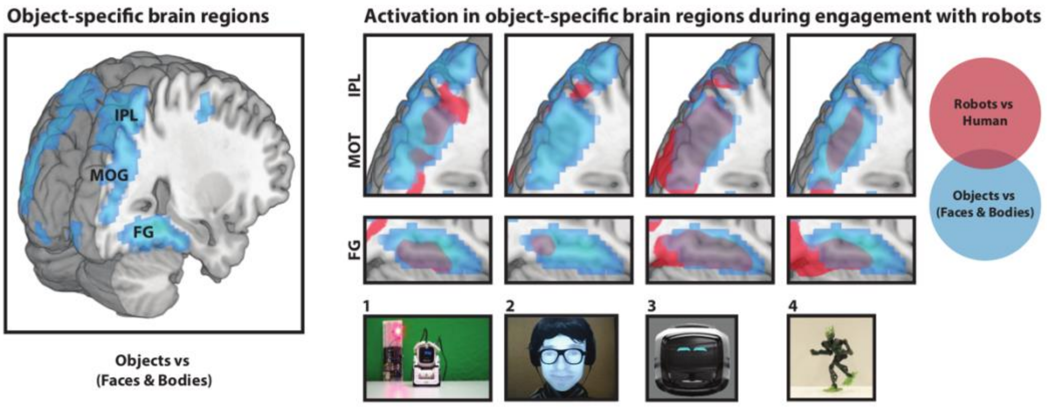

Object-specific brain regions
IPL
MOG
FG
Objects vs
(Faces & Bodies)
Activation in object-specific brain regions during engagement with robots
IPL
MOT
FG
Robots vs
Human
Objects vs
(Faces & Bodies)
1
2
3
4

Figure 13 - Activity in object-specific brain regions during human-robot interactions.

Across several studies that employed different robotic platforms and experimental procedures, a consistent finding is that engaging with robots, compared to engaging with humans, robustly activates object-specific brain regions. (1) Observing robots compared to humans ostensibly experiencing pain or pleasure elicited more activity in the fusiform gyrus (FG), middle occipital gyrus (MOG), and the inferior parietal lobule (IPL) (Cross et al., 2019b). While (2) live interactions with a robot elicited some of these regions (Rauchbauer et al., 2019), observations of (3) emotions and intentions expressed by a robot (Hortensius & Cross, unpublished data), and (4) robotic movements (Cross et al., 2012) lead to widespread activity across these regions. These results indicate the importance of considering brain regions that are selective for object perception. Maps for each study are overlaid on top of an independent object localizer (Pitcher et al., 2011). Unthresholded group-maps are shown for the four studies, while the objects vs faces and bodies statistical map (*n* = 28) for the object localiser is thresholded at FWE < .05 (k = 10). Data for (1) and (2) are from

https://identifiers.org/neurovault.image:108836 and https://identifiers.org/neurovault.image:112530 respectively (Gorgolewski et al., 2015).

Researchers have almost exclusively tested whether robots elicit human-like responses (i.e., do we perceive and react to emotions expressed by a robot similarly to those expressed by a human?). Focusing on direct comparisons between robots and humans does not acknowledge the possibility that robots could elicit subthreshold brain responses in relation to a particular object category. Increased activity in response to human stimuli could therefore be the result of a narrow (univariate) comparison between the two agent categories. A central question in human-robot interaction studies should be what the appropriate comparison categories are for different types of robots. Of course, these could range from humans to objects to animals, and the best answer will naturally depend on the specific research question being tested (Collins, 2019). To establish the place robots might occupy in our social milieu, we need to measure the (dis)similarity to animate agents (e.g., a human or pet) as well as objects (e.g., a phone). Answers to these questions will not only advance our understanding of how people perceive robots and the development of psychological benchmarks for the success of social robots, but also touch upon philosophy, cognitive science and law, which have important implications for society at large (e.g. morality, ethics; Bigman, 2019; Kahn et al., 2006; Prescott, 2017).

5.5. Towards Understanding Real Interactions with Social Robots

Screen-based experiments, third-person observation and one-off or short-term interactions with robots already provide crucial insights on the social cognitive processes that underlie engagement with these novel agents. For the field to move forward, future studies should investigate real and long-term interactions with embodied robots in ecologically valid settings. These studies will provide much needed evidence as to how the human brain negotiates interactions with these agents in real-life settings. Interactions in social spaces that go beyond the laboratory and are relevant to the robotic platform and the user (e.g. schools, care facilities, hospitals) will be particularly important (Broadbent, 2017). The field of social robotics has a long tradition of usability and user experience studies and these investigations will benefit from the sharpened focus on rigor and reproducibility that contemporary psychology and neuroscience bring to the table (**Box II**).

The field of social neuroscience in general still needs to answer the call for taking into account the importance of the second person in an interaction (Schilbach, 2012); this challenge is especially relevant for the study of human interactions with social robots. Paradigms employing free-flow interactions, wherein a recursive perception-action loop exists between two or more agents, are needed. Fortunately, several studies have begun to look at the impact of exposure to or interactions with robots – covering a wider variety of robot design and morphology (Özdem et al., 2016; Wiese et al., 2017). This work is starting to explore neurocognitive aspects of human-robot interactions by integrating information derived from behaviour (e.g., speech, eye gaze) and physiology (e.g., respiration, neural activity) (Cross, Riddoch, et al., 2019; Rauchbauer et al., 2019). One of next steps towards measuring truly unrestricted social interactions is through the use of mobile functional near-infrared spectroscopy (as highlighted below). Combining these state-of-the-art neuroscience methods with new developments in **natural language processing** should enable researchers to step away from **Wizard-of-Oz methods** and provide new ways to examine the social nature of human-robot interactions.

Human-robot interactions are shaped by prior experiences, expectations and beliefs that are continuously updated (Hortensius & Cross, 2018a). It is therefore critical to go beyond contrasting pre- versus post-interaction measures and incorporate longitudinal experimental designs to address questions on experience-dependent plasticity of human social cognition when interacting with social robots. Of note, several commercially available robots allow researchers to collect large datasets per experimental subject over long periods of time, somewhat akin to the experience sampling method (an intensive longitudinal collection of self-report measures). For example, the Cozmo robot (Ciardo et al., 2020; Cross, Riddoch, et al., 2019) collects a rich set of data spanning facial recognition, game performance, and “emotional responses” performed by the robot. Of course, these procedures must consider privacy, data protection and other ethical issues (Rafaeli et al., 2019), but nonetheless offer promise if employed responsibly.

A consideration to keep in mind in the context of social cognition when interacting with robots is the target population that the robots are designed for, and the purpose of these interactions. Whereas two key target populations for robotics developers are children and older adults, participant samples in neuroscience and psychology are predominantly comprise young adults and are often biased towards specific sectors of society (e.g., educated and a relatively high socio-economic status; Henrich et al., 2010). Further, cultural variation exists in the acceptance and uptake of robots (Broadbent, 2017), and this cultural heterogeneity is not fully represented in basic research, which tends to be conducted in industrialized countries, often in western ones. As research on human-robot interaction gradually moves towards broader geographical and societal representation, it is important to consider differences in expectations, attitudes, and beliefs, as well as in prior experiences with robots. This variation needs to be considered in the forms of individual differences (e.g. in learning and plasticity), as well as differences between age groups (e.g. Kirsch & Cross, 2018) and cultures. As one example, one needs to take into consideration that countries such as Japan and South Korea have a longer tradition of research and development in this area (Cameron et al., 2017; Hinz et al., 2019; Jairo Perez-Osorio et al., 2019). Similar to an individualised approach that many technology companies adopt (e.g., social media, streaming services), for which cognitive neuroscience has also advocated (Gordon, 2017), the time is ripe for research into human-robot interaction to adopt methods that are sensitive to and capitalise upon individual differences. Considering how quickly people adopt and can adapt to new technologies, as well as the impact of potential generational differences on attitudes towards such technologies, and the continuous development of new social robotics platforms, it is imperative to keep in mind what a fast-moving and continuously evolving target human-robot interaction is. In order for research in this dynamic area to maximise relevance and generalisability, we argue for the use of specialised methods that enable researchers to map this variation. Combining real and extended interactions with continuous data collection, neuroscience methods and machine learning, could thus be major step towards personalised human-robot interaction (Clabaugh & Matarić, 2018).

5.6. The Promises and Pitfalls of Using Mobile Brain Imaging in Embodied Human-Robot Interaction Studies

New developments in mobile neuroimaging techniques provide the necessary testing ground for how robots might resonate at the social level. One promising technique for studying human-robot interactions is functional near infrared spectroscopy (fNIRS). This technique has been advancing steadily since a connection between human brain function and corresponding light absorption was originally established (Chance et al., 1993). This imaging modality, like fMRI, maps the blood oxygen level dependent response, taking advantage of the transparency of biological tissue (such as skin and bone) in the near-infrared spectrum (for a comprehensive review see Pinti, Tachtsidis, et al., 2020). Light shone on the head with laser diodes or LEDs travels through the skull, scatters back in a banana-shaped curve and is eventually picked up by a detector located at approximately 3 cm separation. The constraints of fNIRS relate to its relatively shallow penetration depth (reaching the outer layers of the cerebral cortex) and relatively low spatial (2-3cm) and temporal resolution (up to 10Hz). It has a lower spatial resolution than fMRI and a slower temporal resolution than EEG, yet brings the advantages of being cost effective, portable and relatively robust to movement artefacts.

These advantages allow for mobile and unobtrusive neuroimaging, thus presenting fNIRS as an optimal candidate for conducting embodied human-robot experiments - especially with under-represented groups such as young children, patients and older adults that often cannot participate in more constraining types of data collection. Researchers in human-robot interaction have embraced fNIRS as a tool to construct feedback loops to control robotic movement or behaviour (Solovey et al., 2012) and as an implicit response evaluation to various robotic systems (Kawaguchi et al., 2012; Mehta, 2019; Strait & Scheutz, 2014b; for a review see Canning & Scheutz, 2013). Various high-quality, commercial imaging systems that allow high-density channel and hyper-scanning set-ups with great potential for research on dyads or groups interacting with a robot are now available, and some recent proof of concept studies have shown the possibility of using fNIRS for connectivity analysis (Bulgarelli et al., 2018).

The transition from lab-constrained experiments that employ screen-based evaluations of social robots to the measurement of unrestricted real-world interactions with physically embodied robots using fNIRS should be a gradual process, adding complexity in a stepwise fashion (Figure 14). For example, in recent years, the brain networks involved in observing social interaction have been mapped in detail (Quadflieg & Koldewyn, 2017). Two regions, the posterior STS and the TPJ, code different aspects of observed interactions (Isik et al., 2017; Walbrin et al., 2018; Walbrin & Koldewyn, 2019). A logical next question is the extent to which the presence and content of interactions with robots is also coded in these regions in third-person encounters. Following on, insights gained from these experiments will pave the way for an embodied research approach where brain activity can be measured during real interactions between humans and robots in unconstrained interactions. In a recent study, for instance, the authors used a GLM-based analysis to automatically identify functional events in fNIRS data, and employed a “brain-first” approach, where instead of being constrained by a block- or event-related task design, a more ecologically valid setting can be chosen (Pinti, 2017). One can envision applying similar methodologies in the context of human-robot interaction experiments.

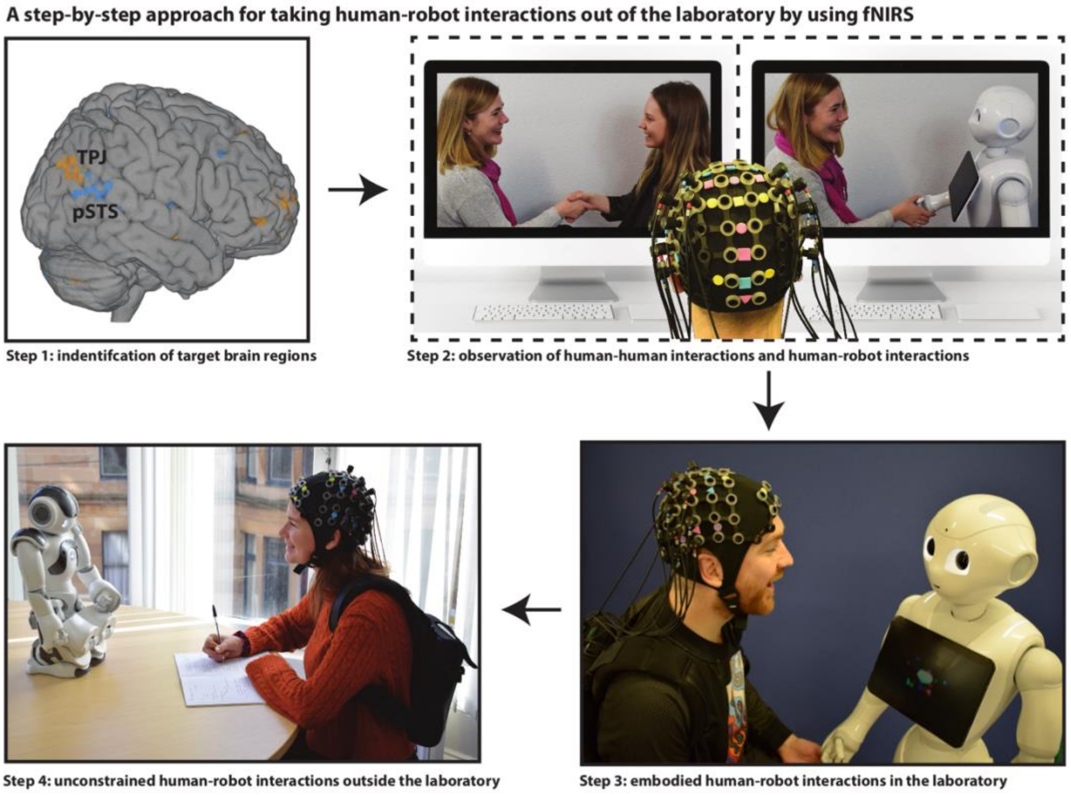

A step-by-step approach for taking human-robot interactions out of the laboratory by using fNIRS
TPJ
pSTS
Step 1: indentifcation of target brain regions
Step 2: observation of human-human interactions and human-robot interactions
Step 4: unconstrained human-robot interactions outside the laboratory
Step 3: embodied human-robot interactions in the laboratory

Figure 14 - Employing functional near infrared spectroscopy for unconstrained human-robot interactions.

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

When using fNIRS in embodied interaction experiments with social robots, several decisions need to be taken: will the device be used to control the robot or inform the evaluation of the robot? How long and “natural” or unconstrained can the interaction be and still yield reliable and interpretable data? Most fNIRS systems, while lightweight and portable in a fitted backpack, cannot be worn for longer than about 45 minutes, due to the pressure of the optodes on participants’ scalps. When performing games or tasks that involve joint movement, another important limitation to keep in mind is that most commercially available social robots are not capable of repeating the same motions for hours on end, as motors can overheat, and batteries run out.

However, despite these constraints, using fNIRS in embodied social robotics studies promises to take us one step closer to following the tenets of a two-person neuroscience (Schilbach, 2012): only by freeing the robots from the screen can we begin to understand how embodied interactions affect cognitive processes in socially relevant areas of the cortex – including the superior temporal sulcus, temporo-parietal cortex and orbitofrontal cortex.

5.7. Concluding remarks

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

5.8. Box 1. Delineating the neural mechanisms of human-robot interaction

How can we examine the functional and temporal changes in neural representations of social cognition during human-robot interaction? Neuroimaging techniques such as EEG and fMRI provide detailed temporal and spatial information on these changes. Traditionally, researchers have looked at relative differences in measures of neural activity during the perception of human and robotic agents. Most research used univariate analyses thereby focussing on distinct networks in the brain, such as the Action Observation Network, Person Perception Network and Theory-of-Mind network. This approach allows researchers to answer questions such as whether brain activation when observing a “happy” robot is higher or lower compared to observing a happy human. In recent years, however, the development and employment of increasingly more detailed analyses, ranging from repetition suppression, to representational similarity analysis, to multi-voxel pattern analysis, provide further and new ways to address questions regarding the overlap of neural architectures for social engagement with humans compared to robots. Repetition suppression enables mapping of potential overlap between similar or dissimilar categories, as repeated stimuli lead to deactivation of regions responsive to these stimuli. For example, does a “happy” robot followed by a happy human (or vice versa) lead to reduced neural activity in a particular region of interest? The presence of repetition suppression would argue for shared neural resources underlying the processing of perceived robotic and human happiness. The critical next step to capture the changes in the representation of social cognition during perception and interaction with social robots is the use of multivariate analyses. Representational similarity analyses can establish the similarity in neural activation during the observation of a happy or angry human and a happy- or angry-appearing robot (Figure 15-A). This approach can test if the neural activation represents a particular stimulus dimension. For example, does activity reflect a representation at the level of agent (activity for robots is dissimilar to humans, regardless of expression) or emotion (activity is dissimilar between happy and angry expressions, but similar across humans or robots). Lastly, a promising way to probe the extent to which perceiving and interacting with humans and robots truly share representations at the neural level is to use multivoxel pattern analyses (Figure 15 -**B**). Instead of measuring magnitude changes, this technique assesses patterns of neural activity that are predictive of specific task conditions, i.e. the representation of different emotions. One way to test possible shared representations is to *train* a classifier to distinguish the observation of a robot displaying happiness from a robot displaying anger, and to *test* this classifier to distinguish a human experiencing happiness from experiencing anger. If the human brain represents perceived human and robot emotions similarly, then the decision criteria of the classifier can be used to distinguish these two different categories. Together, these analytical tools provide new vistas on human social cognition during real and long-term interactions with social robots and the representation thereof.

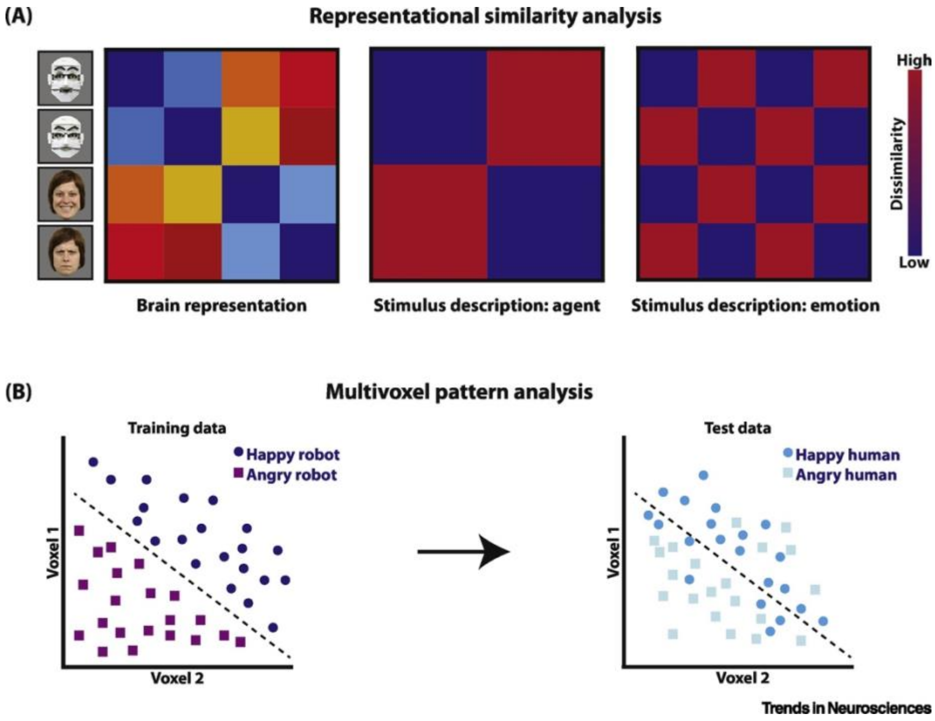


Figure 15 - Towards a Shared Representation of Social Cognition During Human–Robot Interaction.

5.9. Box 2. Integrating Open Science Practices into Human-Robot Interaction studies

The movement towards open science practices and increased reproducibility is gaining momentum across research domains in the life and physical sciences, including psychology and human neuroscience (Munafò, 2017; Poldrack et al., 2019). Similarly, these issues are acknowledged in Artificial Intelligence research (Hutson, 2018), and have recently been further reflected upon by robotics researchers (Bethel & Murphy, 2010; Eyssel, 2017; Irfan, Kennedy, Lemaignan, et al., 2018). Issues of transparency and reproducibility are especially important for investigations of the neurocognitive mechanisms supporting human-robot interaction. Integrating methods and tools from psychology and neuroscience, researchers not only face reproducibility issues key to these fields (e.g. reliability of fMRI findings (Button et al., 2013), and researchers' degrees of freedom in pre-processing pipelines of fNIRS and fMRI data (Carp, 2012; Pinti et al., 2019), but also issues specific to the field of social robotics (e.g. cross-platform generalisability, access to expensive and bespoke robotic platforms). Encouragingly, experimental reform is being implemented in the human-robot interaction community, with the 2020 ACM/ IEEE International Conference on Human-Robot Interaction being the first to invite replication studies for submission. In recent years, psychologists and neuroscientists are more broadly embracing open science practices, which will help to remedy many of the abovementioned issues. Concrete actions along these lines include taking steps like pre-registering studies, conducting replication studies, sharing research materials and (anonymized) data, as well as posting pre-print articles (Munafò, 2017; Poldrack et al., 2019). This scientific reform can especially benefit human-robot interaction research, as studies are often resource- and time-intensive and include relatively small samples of subjects. Sharing data and scripts will enable the wider community to conduct secondary and meta-analyses and exploratory tests on published data. Sharing of research resources and products should also contribute to a more inclusive community, giving, for example, access to data from bespoke robotic platforms. Finally, a movement toward greater openness and transparency should facilitate more exchange between disciplines as well as a more robust human-robot interaction literature, by creating an ecosystem conducive of cross-platform replication. One question the field needs to address is the cross-platform generalizability of previous findings (Cross, Hortensius, et al., 2019; Hortensius et al., 2018b). Developmental social robotics already successfully implements artificial architectures to test cross-platform generalizability (Cangelosi & Schlesinger, 2018) and future research should further incorporate this practice to replicate and extend previous findings. Moving forward, the implementation of open science practices can help facilitate more reproducible user studies and can foster a common ground in terms of methodology between human-robot interaction researchers and cognitive neuroscientists.

5.10. Glossary

Action Observation Network – a collection of brain regions comprising parts of parietal, premotor and occipitotemporal cortices that responds when watching other agents (human or robotic) in action.

Automatic imitation – see **motor interference.**

Brain-computer interface – a setup that allows for signal relay between the brain and an external device, such as robot, usually via a computer.

Gaze cueing paradigm – a commonly used psychological paradigm used to investigate the mechanisms of joint attention. The gaze of an observed other (human or non-human, physically present or viewed on a screen) either looks towards or away from a visual target the participant is required to attend to, and the cost in a participant's response time is thought to be a measure of social engagement.

Human-robot interaction – **see social robotics.**

Mentalising – a cognitive process by which an individual reflects on, explores and interprets their own and others' thoughts and feelings, and how these influence behaviour and actions.

Motor Interference – Observing others perform movements incongruent to one's own has been found to produce motor interference. Motor interference is closely related to automatic imitation, a phenomenon that describes the tendency of humans to implicitly imitate others' actions and other social cues.

Natural language processing – field of study concerned with the recognition and production of natural language by computers and algorithms.

Pain Matrix – collection of brain regions associated with empathy and emotional processing when seeing another individual in pain or distress. Primary nodes of this network include bilateral anterior insular and medial anterior cingulate cortices

Person Perception Network – a collection of brain regions responsive to other individuals, especially their faces and bodies. Regions include the fusiform face area and extrastriate body area, among others.

Prediction Error - a mismatch between a predicted and observed response.

Repetition Suppression – In a brain imaging context, this refers to a reduction in a neural response that emerges when a stimulus (or a certain aspect of a stimulus) is repeated more than once. Also referred to as repetition priming.

Social robotics – this term encompasses a wide variety of research relating to robots designed to engage humans on a social level, often framed in a companionship or assistance context. **Human-robot interaction** is one facet of this diverse field, which specifically investigates how humans perceive and interact with robots.

Social robotics winter –a term used to describe the current disillusionment surrounding social robots, as technological developments have failed to live up to the hopes and expectations fed by robotic depictions in film, television, and other media, as well as the failure of several recent robotics start-ups.

Theory of Mind - the ability to attribute other mental states (thoughts, desires, and intentions) to other individuals. Commonly associated with a network of brain regions, the **Theory-of-Mind network** including the medial prefrontal cortex, bilateral temporoparietal junction and the precuneus.

Uncanny Valley Hypothesis – humans prefer anthropomorphic agents but reject them if they appear too human-like - to what extent the uncanny valley is an artefact of contemporary experimental procedures remains unknown.

Wizard-of-Oz – describes an experimental set-up in which the robot does not operate autonomously, but rather is controlled by the experimenter, thus resembling the trickster turned wizard in the eponymous film.

5.11. Outstanding questions

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

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

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

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

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

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

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

5.12. Acknowledgements

The authors thank Michaela Kent, Te-Yi Hsieh, Laura Jastrzab, Ben Jones, Richard Ramsey, Kohinoor Darda, and gratefully acknowledge funding from the European Research Council (H2020 ERC-2015-StG-67720-SOCIALROBOTS) and Leverhulme Trust to E.S.C., and from the Bial Foundation to R.H.

Chapter 6 Validating functional near-infrared spectroscopy as a tool for studying embodied social interactions with robots

6. Abstract

Functional near infrared spectroscopy (fNIRS) is a promising tool for the evaluation of embodied human-robot encounters. In this study, we compared the quality of the fNIRS signal with the fMRI blood oxygen-level dependent (BOLD) response, to establish the detection rate of brain activity with each modality, using a robust Theory of Mind (ToM) localiser task. Using a new method for the digitisation of fNIRS probe positions, we also investigated overlap sensitivity of the probes and individual participants' functional regions of interest. We found that on the individual subject level, the localiser evoked robust activity in the bilateral temporoparietal junction (TPJ), as measured with fMRI. However, the channel-wise fNIRS analysis showed more variable results. Finally, the photogrammetry and subsequent co-registration with subjects' anatomical brain scans was successfully accomplished for every subject, revealing how inter-individual differences in the subjects' brain anatomy could have contributed to the lower signal-to-noise (SNR) ratio using fNIRS.

6.1. Introduction

Recent advances in fNIRS have allowed researchers to transform brain imaging experiments from those where participants observe social interactions into experiments where participants take part in actual embodied social interactions, moving the field closer towards a more ecologically valid ‘second-person neuroscience’ (Pinti et al., 2018; Schilbach, 2012). There are clear advantages of using fNIRS over other brain imaging methods like electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) in naturalistic HRI environments. Both social neuroscientists and HRI researchers have reflected on the synergies that emerge from implementing fNIRS into evaluations of social robots, i.e. those robots that are designed to engage humans on an interpersonal level (Henschel, Hortensius, et al., 2020b). Indeed, Strait and Scheutz (2014a, p. 1) see fNIRS as an innovative tool, which can “build […] a literal bridge between robotics and neuroscience” either as a brain-computer interface or as an offline evaluation method. In comparison to EEG and fMRI, fNIRS offers a reasonable option for *in-situ* neuroimaging, although with lower spatial resolution than fMRI, and lower temporal resolution than EEG. Thus, HRI researchers seeking to increase the ecological validity of their experiments with fNIRS should keep these limitations in mind.

While EEG offers higher temporal resolution in the millisecond range, it is extremely sensitive to subject motion, and is thus only of limited use in embodied encounters with robots. Functional MRI, which offers relatively superior spatial resolution, also constrains the scope of participants’ motor responses. Young children in particular, who have been shown to exhibit differential reactions to social robots (as compared to adults), are not ideal scanner participants due to the signal’s susceptibility to motion (Greene et al., 2018; Vollmer et al., 2018). Mobile fNIRS systems, on the other hand, are more robust to subject motion, as long as the optodes are tightly fixed to the head of the participant (Huppert et al., 2009). With this brain imaging modality, very young children can move more freely and can be directly supported by their parents while data are recorded (Powell et al., 2018). To summarize, fNIRS might not have the spatial resolution of fMRI or the temporal resolution of EEG, yet it fares better than these more established modalities in terms of robustness to motion, as well as better portability and lower cost associated with scanning time.

In addition to these advantages, fNIRS has also been emphasized as a valuable tool to objectively assess human-robot encounters (Balconi & Molteni, 2016). Wiese and colleagues' (2017) recent perspective on the intersection between neuroscience and HRI critically reflects on current empirical work with social robots, which, they argue, is limited by inconsistent assessments of robots. The authors argue that in order to build successful social robots (where 'successful' means robots that are accepted by humans as a social companion), behavioural and neuroscientific methods are needed to *systematically* probe neurocognitive mechanisms at play (Chevalier et al., 2020; Wiese et al., 2017). Chevalier and colleagues (2020) stress that cognitive neuroscience can bring this meticulous approach to the table when studying human-robot interaction, as specific, isolated (social-)cognitive mechanisms are targeted and observed.

Recently, we proposed that before adapting embodied paradigms for human-robot interaction experiments, an iterative process should occur: first, researchers should identify a target region relevant in human-*human* interaction and replicate activity of this region with fNIRS in screen-based paradigms (Henschel, Hortensius, et al., 2020b). In the next step - if robust activation is found, complexity can be added in the form of embodied interactions with robots in the laboratory. In the final step, brain activity can be recorded in naturalistic environments, for example in care-homes or children's nurseries – current popular use cases for human-robot interaction (Pinti, Devoto, et al., 2020; Pinti et al., 2018; Quaresima & Ferrari, 2019). A rationale for using this gradual approach is that the evaluation of each step can inform the pre-registration of the methodology and analytical plans for the more complex, following steps. Further, this method could also help social neuroscientists uncover discrepancies or overlap between real encounters and their passive observation (Schilbach, 2012).

Following our proposal, in this proof-of-concept study, we sought to test the feasibility of fNIRS as a tool for embodied social interaction experiments with robots. The main focus of the investigation was to transparently report and highlight any challenges we encountered, as well as test a new methodology to spatially register the probes, in order to inform the development of future fNIRS-for-HRI studies (Clausner et al., 2017; Hu et al., 2020).

6.2. Mechanisms of optical brain imaging

fNIRS, an optical imaging technique, benefits from the fact that biological tissue, like skin and bone, is relatively invisible in the near-infrared light range (600-900nm) and thus offers an ‘optical window’ into the activity of the outer layers of the cerebral cortex (Ferrari & Quaresima, 2012; Scholkmann et al., 2014). Oxygenated hemoglobin (HbO) and deoxygenated (HbR) are chromophores - they absorb near-infrared light to a different extent. Based on the differential absorbance of light, concentrations of HbO and HbR can be calculated, which is closely tied to the stimulus-evoked response, eliciting local vascular and metabolic effects that are commonly known as the hemodynamic response function (HRF; Huppert et al., 2009).

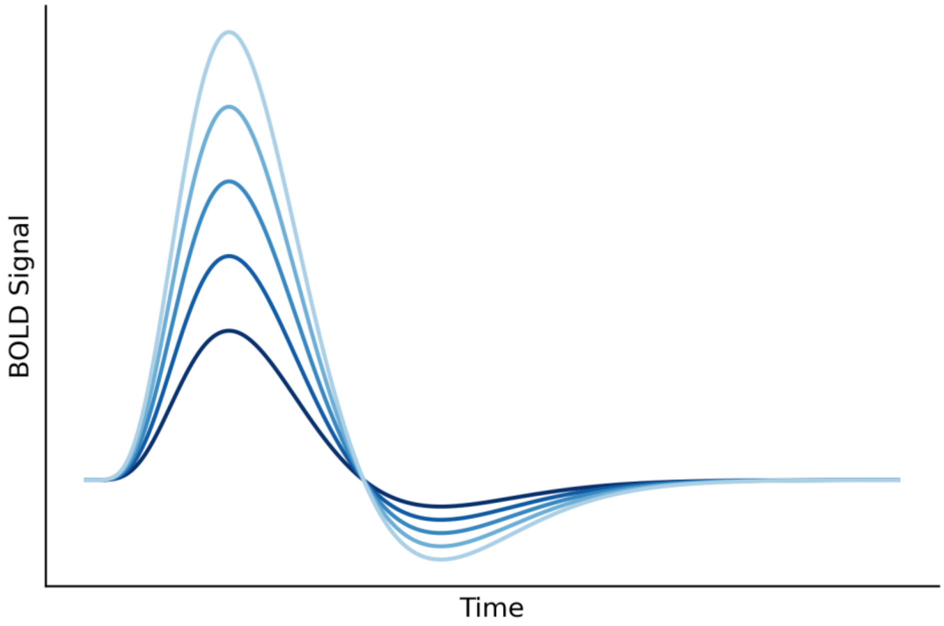


Figure 16 - Schematic depiction of the shape of the hemodynamic response function (HRF)

Here the different traces relate to differences in BOLD response magnitude due to variations in stimulus conditions – or differences in individual participants (see also Poldrack et al, 2011, p. 3). Graph by Dan Gale used with permission.

The HRF maps the time-lagged vascular response induced by brain activity, which is followed by a local oversupply of HbO, coupled with a decrease in HbR.

Thus, similar to fMRI, the fNIRS signal is based on neurovascular coupling (Wan et al., 2006), as it measures the increase in cerebral blood flow following neuronal activation (Figure 16). The HRF is elicited through the presentation of a stimulus, and for healthy adults it peaks at about 5-6 seconds (Friston et al., 2000; Poldrack et al., 2011). Depending on the stimulus (or repeated presentation of the stimulus), the magnitude of the HRF may differ (Figure 16).

Light shone on the head with the light-emitting optodes penetrates a little less than half the source-detector separation, usually spaced at 3cm distance for adults (Pinti et al., 2019). The investigated tissue volume that lies in between is referred to as a 'channel'. fNIRS, despite being constrained to a relatively shallow penetration depth, has been used to study a multitude of processes, including motor, visual, language, auditory and cognitive systems (Huppert et al., 2009).

6.3. Validating fNIRS with fMRI

To validate fNIRS as a feasible tool for embodied HRI studies, we followed the established tradition of fMRI-based verification, as fMRI is still considered the "gold standard for neuroimaging" (Wijeakumar et al., 2017, p. 204). Using the new spatial registration method, it was our goal to establish an anatomical and functional ground truth using MRI and relate this ground truth back to the fNIRS probe placement. Evidence from previous studies validating NIRS with fMRI has shown that overall there is a relatively strong degree of correspondence between the signals. Cui and colleagues (2011) scanned participants concurrently with fMRI and fNIRS, placing optodes over frontal and parietal regions and testing a battery of cognitive tasks. The fNIRS measure showed a lower signal-to-noise (SNR) ratio than fMRI (as would be expected). However, the authors reported a significant correlation between HbO and the BOLD response, which they attribute to the fact that HbO has a higher SNR than HbR. Task-characteristics also played a role: motor, language and visual tasks showed robust correlations, however, the authors also report that longer scalp-brain distance negatively impacted the signal correlations. A mean scalp-brain distance of 16.8mm was reported, importantly, with a larger distance in parietal regions (i.e. those of interest to social neuroscientists and HRI researchers). The authors explicitly encouraged future research to investigate the correspondence of fNIRS and fMRI for event-related (rather than block-design) tasks with more subtle effects, which we addressed in the current validation study.

Noah and colleagues (2015) compared fNIRS and fMRI by sequentially recording brain activity using an adapted dance video game task designed to assess the integration of multi-modal stimuli. The authors correlated the two signals and found good correspondence between them, concluding that this constitutes positive evidence for the feasibility of replicating fMRI findings using fNIRS in a naturalistic scenario. In their scanner-adapted version of *Dance Dance Revolution*, participants responded on a modified foot platform in a block-design task consisting of rest and play epochs. These authors highlight that they chose a block design paradigm to “yield strong cortical responses in both procedures” (p.9). In a sophisticated study by Wijeakumar and team (2017), fNIRS and fMRI was recorded simultaneously, and to test the correspondence of the signals, an image-reconstruction approach for the fNIRS data was used. To allow for direct comparison with the voxel-based fMRI results, the authors transformed the data from channel to voxel space. Here they found significant voxel-wise correlations for all experimental conditions in frontoparietal and temporal cortices.

Overall, the reviewed studies seem to generally support good methodological correspondence between fMRI and fNIRS, with some variability in the correlations that can be attributed to larger scalp-brain-distances in fNIRS.

6.4. The role of the TPJ in HRI

In the current study, we targeted a hub-region central to day-to-day social cognition – i.e. “a suitable starting point” when transitioning to the investigation of embodied human-robot encounters (Henschel, Hortensius, et al., 2020b, p. 5). During human-*human* interaction, the TPJ forms part of a domain-general network, involved in person perception, action observation and mentalizing, but also non-social functions (Darda et al., 2018; Olmen, 2018; Schurz et al., 2014, 2017). Converging results from fMRI and fNIRS studies have added further support for TPJ playing a role in imitation and self-other distinction in some observation-only and embodied human-human interaction paradigms (Gallagher & Frith, 2003; Oliver et al., 2018; Olmen, 2018; Schurz et al., 2014).

Due to its crucial role in controlling social interactions with other humans, the TPJ has been a central subject of investigation in HRI brain imaging studies as well (Hortensius & Cross, 2018a). In an fMRI study by Gobbini and colleagues (2011), participants viewed emotionally evocative facial expressions by humans and robots. The researchers found that activity in regions commonly referred to as the ToM network – which are engaged when we are thinking about the inner mental lives of others – was reduced. Specifically, participants' right medial prefrontal cortex (MPFC) and right TPJ showed less activation in the robot condition. Hortensius and Cross (2018a) reviewed several fMRI studies reporting reduced engagement of the bilateral TPJ when robots are observed, compared to cues to humanness, and summarised that this reduced activation can be linked not so much to the external features of robots, but rather beliefs and expectations about them. Hence, an interesting question for future embodied HRI studies is whether this reduced activation of the mentalising regions of the brain holds in direct interactions with autonomous or Wizard-of-Oz controlled robots.

To effectively target the bilateral TPJ, a validated functional localiser task from the fMRI literature was selected for the present study (Jacoby et al., 2016; Richardson et al., 2018; Richardson & Saxe, 2020). In their original study, Jacoby and colleagues found that the functional region of interest (fROI) detection rate for this short Pixar-animated movie was high for the bilateral TPJ: for the left TPJ (LTPJ), it was successful for 17 out of 17 participants, and for the right TPJ (RTPJ), it was able to elicit functional activity in 16 out of the 17 participants. Since then, many fMRI studies have enthusiastically implemented the localizer, and a follow-up experiment by Richardson and colleagues (2018) showed that ToM regions, including the bilateral TPJ, are developed and functionally distinct very early in life – as early as 3.5 years of age. The rationale for using this short functional localiser task was that it has shown to be efficient in eliciting robust activity on the single-subject level. Using the fMRI activity as our ground truth, we wanted to investigate the translation of these findings into the fNIRS modality.

6.5. The current study: research questions and hypotheses

The main research question we wanted to address in the current study relates to overlap sensitivity, i.e. how much would the spatially registered fNIRS probes overlap with the functional MRI activity on the single subject level? Related to this, we asked how consistent the optodes placements would be over our region of interest: bilateral TPJ. The second major question we wanted to address was how the detection rate for functional activity on the single-subject level would compare across the two modalities.

As this was an exploratory, proof-of-concept study, we had no directional hypotheses about either of these two research questions, however, for each one of the two modalities we expected to see increased activity in the bilateral TPJ for the contrast of interest: mentalising events in the movie task versus pain events.

6.6. Methods

6.6.1. Open science statement

Owing to the exploratory nature of this study, no pre-registration was published before data collection, however, the insights gained in here will guide future pre-registrations. The task was presented as part of a larger validation project posing additional questions regarding test-retest replicability (for both modalities) and spatial specificity of the fNIRS modality. Hence, here we report only the measures, data exclusions and the sample size determination rule for a subsection of this wider project (BOLDlight): the bilateral TPJ functional localiser task. Brain and the spatial registration data will be shared at Neurovault (https://neurovault.org/) after they are appropriately anonymized and de-faced. The data sharing issue is especially salient for the sensitive data recorded during the photogrammetry-based spatial registration. Homölle and Ooostenveld (2019) have recommended that only the X,Y and Z coordinates of the final probe positions should be shared as point clouds, to avoid the identification of subject identities.

6.6.2. Participants

Matching the sample sizes of previous fMRI-fNIRS validation studies (Cui et al., 2011; Wijeakumar et al., 2017), twelve subjects (26.4 ± 6.4 years of age, 8 female/ 4 male), who met our selection criteria of normal or corrected-to-normal vision, no learning disability and normal hearing abilities were invited to participate for monetary compensation. Subjects were contacted by the investigators prior to the testing date to ensure that they understood what their participation entailed, as well as to rule out any contraindications for the fMRI aspect of the experiment. All procedures were approved by the University of Glasgow local ethics committee (Ethics numbers: 300180151 and 300180301) and the subjects provided written informed consent.

6.6.3. Experimental Procedure

Participants passively observed the CGI-animated short film "Partly Cloudy" (Pixar Animation Studios). The movie played after a 10 second fixation cross and was presented in PsychoPy (Peirce, 2007). Subjects were given the instruction to quietly sit (or lie, in fMRI) and observe the movie, while making as few movements as possible. In the 5.36 minutes film, clouds conjure human and animal babies, which are delivered to earth by helpful storks (Figure 17). Throughout the movie, painful events take place (a stork is injured by the spikes of a baby porcupine), as well as ToM, during which the observer is prompted to consider a character's thoughts (for example: the stork, who has been repeatedly injured by the dangerous babies is caught looking at a cloud conjuring fluffy puppies). In the original validation study by Jacoby and colleagues (2016), 4 types of events were coded ('control', 'social', 'pain' and 'mental'), however, in this study we selected those 'pain' and 'mental' events identified in the reverse correlation analysis by Richardson and colleagues (2018).

PIXAR ANIMATION STUDIOS
Partly Cloudy | Publicity Image | Pixar Creative Services
generated from element: cs_comp
c1_3apub.pub8.210.tif - 2009:04:15 11:30:24 - (1920 x 1080)

Figure 17 - Screenshot taken from the Partly Cloudy short film.
Image by Plidezus Leo on Flickr (CC BY-NC-SA 2.0)

We selected those events that reliably replicated in an independent sample of adults, which resulted in 7 ToM events and 9 pain events, with durations ranging from 4 to 16 seconds. The order of the scanning modalities (fMRI or fNIRS) was counterbalanced – half of the participants saw the movie first in the MRI scanner, half in the fNIRS lab. In the fNIRS modality, we encountered a problem with manually sending the triggers – sometimes when the trigger was elicited, the task did not start right away. In those cases, a second trigger was sent, which was the one we used for the analysis. However, in all three cases there was only a 100-millisecond difference between the first and second trigger.

6.6.4. Photogrammetry-based spatial registration

A particular challenge associated with fNIRS is the consistent placement of optodes onto participants' heads, once a target brain area has been identified (Powell et al., 2018). The standard approach is to follow the landmarks (nasion, inion, auricles and Cz) associated with the 10-20 system and spatially register optode locations with a 3D magnetic digitizer (such as the Polhemus) to either subject brain anatomy or an age-matched template (Clausner et al., 2017; Noah et al., 2015). However, several limitations have been identified that are associated with this method: the high cost of such electromagnetic digitizers (the Polhemus costs ~ $8000), distortions introduced by nearby metal objects, and low spatial accuracy (Clausner et al., 2017; Homölle & Oostenveld, 2019). To address these issues, recent advances in computer-vision technology have allowed for a more accurate and cost-effective registration method: photogrammetry, which describes the method of building 3D models based on 2D images, reconstructed from information of overlapping pictures (Wesencraft & Clancy, 2019). This new method of spatial registration, which was developed by Clausner and colleagues (2017) for EEG electrodes and MEG fiducial markers, seems to promise higher accuracy and flexibility.

The authors developed an open-source toolbox for Matlab (janus3D), which maps the spatial location of electrodes to a participants' anatomical brain scan via a matching technique that takes advantage of rigid facial features. Using a replica adult head, Clausner and colleagues (2017) compared the performance of this technique to the performance of the Polhemus magnetic digitizer and found that EEG electrodes were co-registered with an average error of less than .10mm with photogrammetry, while the electromagnetic digitizer resulted in an average error of 6.1mm. However, the success of the 3D model reconstruction with a participant wearing the EEG cap depends on colour difference information, which raises the question how successful the translation of this technique will be when constructing 3D models of participants wearing fNIRS caps.

An initial validation study by Hu and colleagues (2020), who were also interested in implementing Clausner and colleagues' toolbox for fNIRS, gives tentative support for the feasibility of the method. Although the authors reported a larger registration error than Clausner and colleagues, they compared the technique to spatial registration of fNIRS optodes with an MRI-derived spatial registration technique (using vitamin E capsules sewn in the fNIRS cap). The authors attribute the resulting larger error to the movement of the fNIRS cap when participants were placed in the MRI head-coil. Complementing the findings of Hu and colleagues (2020), in the current validation study, we tested the feasibility of the new photogrammetry-based spatial registration method for fNIRS.

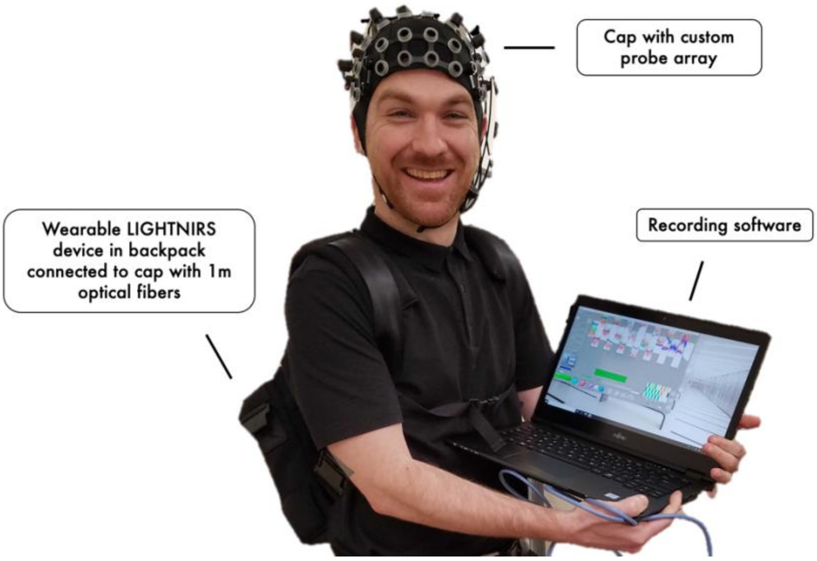

Cap with custom probe array
Wearable LIGHTNIRS device in backpack connected to cap with 1m optical fibers
Recording software

Figure 18 - The Shimadzu LIGHTNIRS system demonstrated in its portable configuration.

During the experiment, subjects were seated and the fNIRS recording device was placed on a table behind them. The probes were protected from stray light with a dark silk cap placed on top of the cap. The subject gave consent for this image to be shared.

To digitise the optode positions, we followed the photogrammetry for EEG procedure described by Clausner and colleagues (2017). To ensure consistency, the Shimadzu cap (Figure 18) was placed on participants' heads according to the 10-20 landmarks: nasion, inion and left and right preauricular points (Jasper, 1958). Additional colourful stickers (Figure 19) were added to the cap to aid the 3D reconstruction of the head models, which as described above, relies on colour difference information (Clausner et al., 2017). The centre of the cap was aligned with the centre of the head, as measured by the distance between nasion and inion, as well as the left and right preauricular points. However, due to constraints of the setup (i.e. challenges of precisely replicating placement between participants), for a subset of the subjects the landmark points themselves were not marked and spatially registered.

In the first step of the spatial registration, pictures were taken with a Canon D3500 DSLR camera while participants rotated on the chair in front of a chroma key green screen. To ensure consistent lighting, two studio lights were placed to the left and right of the subject (Figure 19). Participants rotated with closed eyes (to ensure minimal facial motion) in small steps of 10 degrees for three different height settings. The procedure was repeated twice, to allow for back-up photos, should the first run not result in a good quality model. Between 40 and 90 pictures were taken for each participant on each of the two photogrammetry runs. The photography was not timed; however, this step usually took between 15 and 25 minutes to complete.

The configuration of the camera and further information are described in more detail in our manual (Henschel, Kent, et al., 2020c), which can be found at [https://zenodo.org/record/4146985#.X5ny0VngphE]. The processing of the images was performed offline, after data collection was completed. To build the 3D head-model, images were first loaded into janus3D, and using the Photo Masker functionality, a mask of the chroma green background was created for each picture. The pictures and their corresponding masks were then loaded into Metashape (Agisoft).

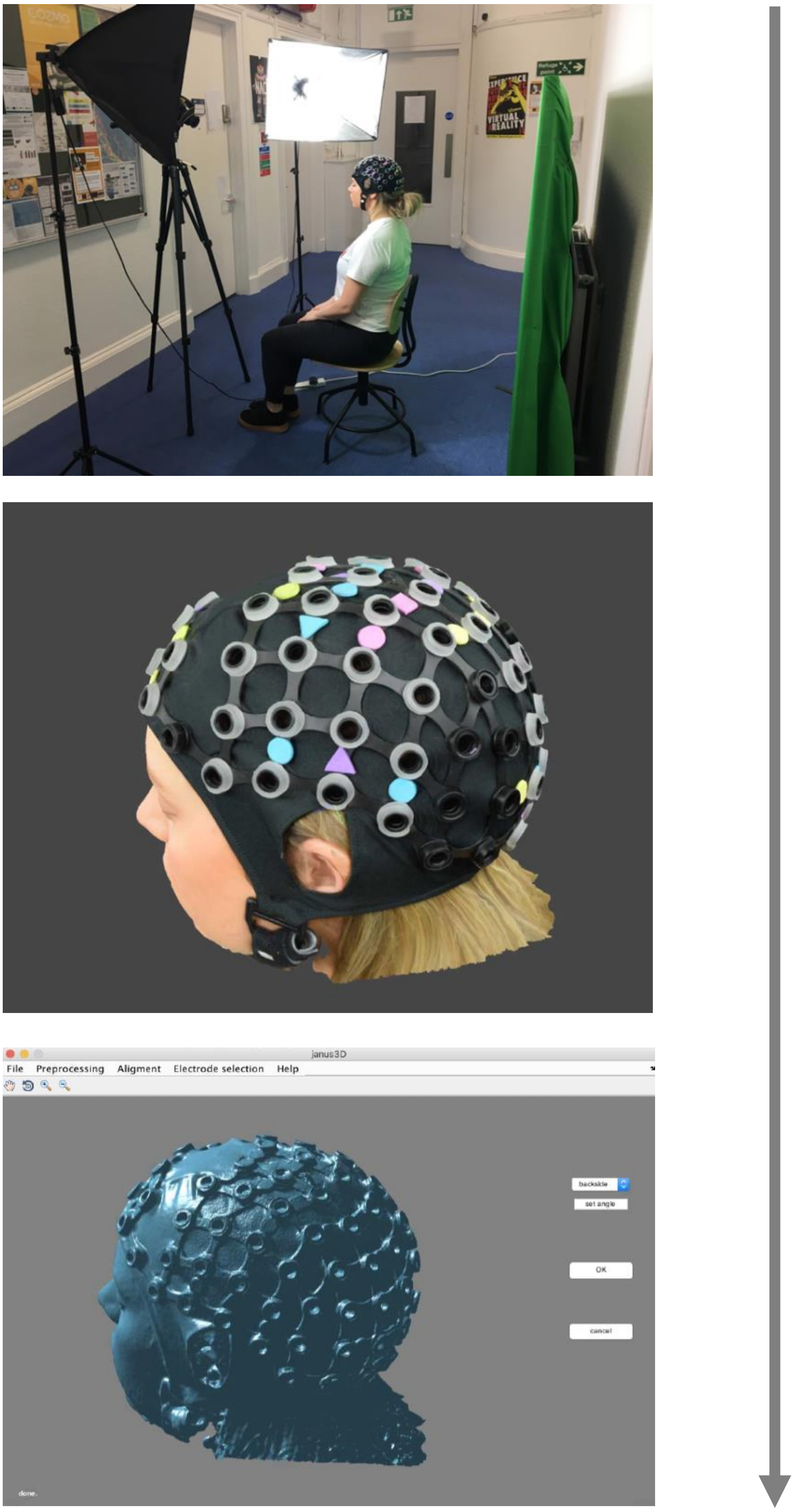


Figure 19 - Overview of the photogrammetry-based optode digitisation.

Three steps are involved in the photogrammetry-based co-registration of the optodes: taking the photos, building the 3D head model in Metashape and finally co-registering the model and the subject's MRI anatomical scan in janus3D. The subject gave consent for these images to be shared.

In Metashape, a matching point cloud was built, then a dense point cloud, a polygonal mesh and finally a dense mesh. To obtain the final 3D head model, texture information was added. The resulting object file, along with participants MRI anatomical scan, was then loaded in janus3D, to perform the final step of registering the optode positions to subject brain anatomy (Figure 19). For 11 out of 12 subjects we obtained good quality 3D head models. However, even the head model that was classified as subpar allowed us to visually identify and mark optode positions. Between the two runs, the best quality head model was selected for co-registration in janus3D.

In the first step of the co-registration, the head models are rotated from Metashape's arbitrary to MRI space. Then the head model and MRI scan are aligned by outlining the side profile of the subject's face. Next, the facial features are matched to the anatomical scan. The resulting overlay had to be manually corrected for most subjects. Three out of the 12 subjects already had de-faced MRI anatomical scans. For these three cases, the manual correction of the alignment was more comprehensive. After manual correction of the alignment, the optode holders of interest were each manually marked, as the automatic selection algorithm relies on the contrast between the "electrodes and surrounding texture" (Clausner et al., 2017, p. 6). In the case of the grey optode holders and the black cap, the contrast was not salient enough for the automatic detection to reliably work.

Extraction of the optode coordinates from the matrix file was completed using the MarsBaR toolbox (Brett et al., 2002) in SPM12 (Wellcome Trust Centre for Neuroimaging, UCL, London) in MATLAB version 8.5 (The MathWorks Inc., Natick, USA, 2018b). Using MarsBaR, it is possible to create spheres with the exact coordinates obtained during co-registration. This step created 16 nifti files (from 8 source and 8 detector locations), each one corresponding to the location of one of the optodes. To combine these nifti files, again MarsBaR was utilised to create a "master" nifti file containing the 16 spheres.

Finally, the fNIRS optode coordinates were mapped onto the anatomical images of the brain in MRIcro (Rorden, 2007). This step was completed in native space for each participant.

6.7. Data acquisition (fNIRS)

We used a wearable Shimadzu LIGHTNIRS system (Kyoto, Japan) to record HbO, HbR and total hemoglobin concentration changes. 8 light sources (near-infrared semiconductor lasers) and 8 detectors (avalanche photodiodes, APD) were arranged covering participants' putative temporoparietal cortex following a 2x4 (R), 2x4 (L) probe geometry (Figure 20). This probe geometry approximately matched previous fNIRS studies' probe geometry which targeted the TPJ (Hyde et al., 2015; Oliver et al., 2018; Olmen, 2018). The configuration resulted in 20 logical channels: #1-10 on the right side of the head, #11-20 on the left side. The LIGHTNIRS uses 3 wavelengths to account for scattering when converting changes in optical density to HbO and HbR concentrations: 780, 805, and 830nm (Pinti et al., 2018). Before the optodes were attached, hair under the cap probe holders was removed with blunt knitting needles, as other fNIRS researchers have highlighted that removing the hair from the optode surface is one of the key strategies to ensure a good signal to noise ratio (Noah et al., 2015).

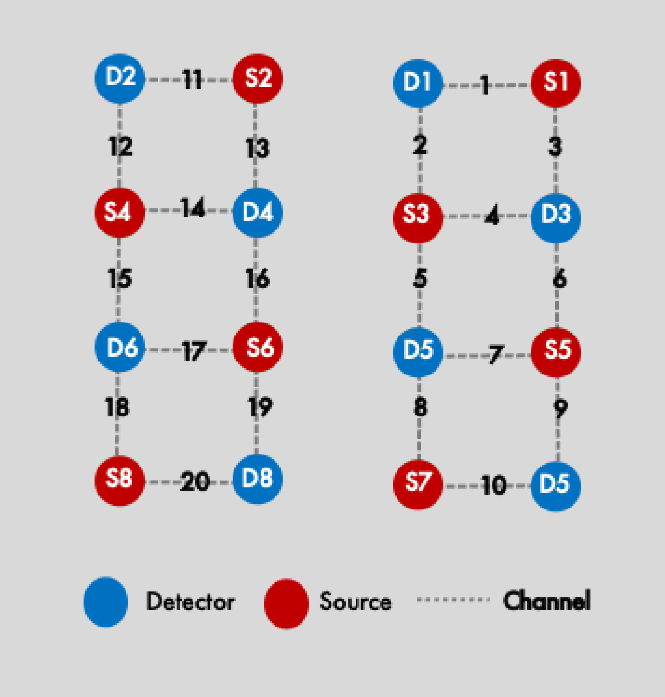


Figure 20 - Schematic representation of the fNIRS probe geometry. The design of this figure is modelled after Pinti et al. (2020a).

Once the optodes were connected to the cap, a probe check, i.e. the assessment of the quality of optical coupling, in the Shimadzu "fNIRS" software was conducted. Channels with low SNR or with an overflow error were adjusted on the participant's head until they showed satisfactory signal quality. Further SNR improvements were undertaken by adjusting the voltage of the APDs.

6.8. Data analysis (fNIRS)

Data was recorded with a 13.33Hz sampling frequency. The fNIRS data was preprocessed with the open-source Matlab toolbox Homer2 (Huppert et al., 2009). Our preprocessing pipeline was derived from the one described by Pinti and colleagues (2020a), taking into consideration recent recommendations for standardizing NIRS data processing pipelines (Pinti et al., 2019). Prior to loading the files into Homer2, the raw intensity data was transformed from the proprietary Shimadzu format into the nirs file format with the help of a custom Matlab script.

Raw intensity was processed with the *enPruneChannels* function, which automatically removes channels from the measurement if the signal is too strong, too weak or the standard deviation is too great (Perry, 2019; Powell et al., 2018). One channel (channel 3, subject 11) was discarded. Given the fact that participants were instructed to sit still and observe the movie, this low exclusion rate is perhaps not surprising. The raw intensity data was then transformed into changes in optical density (function, *hmrIntensity2OD*). We then removed motion artifacts using a wavelet-based approach (function, *hmrMotionCorrectWavelet*, iqr=1.5), which has been shown to be the most effective strategy for identifying spike artifacts elicited by decoupling of optical probes from the skin (Molavi & Dumont, 2012; Pinti et al., 2018). Next, many sources of noise (heart rate, low frequency noise, & slow trends) were removed using a standard third-order Butterworth bandpass filter (function, *hmrBandpassFilt*, band-pass frequency range [0.01, 0.4]). The optical density rather than the concentration signal was filtered to avoid "artifact contaminated data in calculation of oxygenated and deoxygenated hemoglobin" (Molavi & Dumont, 2012, p. 263). Then, using the modified Beer-Lambert law, the changes in optical density were converted to changes in concentration (function, *hmrOD2Conc*) with a differential path length factor of 6.

One chromophore (HbO) was considered for further analysis, as past work has focused on this signal due to a higher SNR and better correspondence with the BOLD response - however, at the same time this signal may be more confounded by physiological noise (Cui et al., 2011; Hyde et al., 2015; Pinti, Tachtsidis, et al., 2020b). A general-linear modelling approach was chosen, as this takes advantage of the fast event-related task design and is considered more powerful than block averaging (Pinti et al., 2019; Wijeakumar et al., 2017). The design matrix was composed of the two-task related regressors (pain events and mentalizing events), as well as the constant term. We investigated the Mental>Pain contrast of interest. Beta values were estimated for each channel, for each participant. One-sample *t*-tests were conducted to investigate the hypothesis that the signal in one channel was active at a significance level of *a* =.05. The *p*-values in this exploratory, single-subject level analysis were not corrected for multiple comparisons.

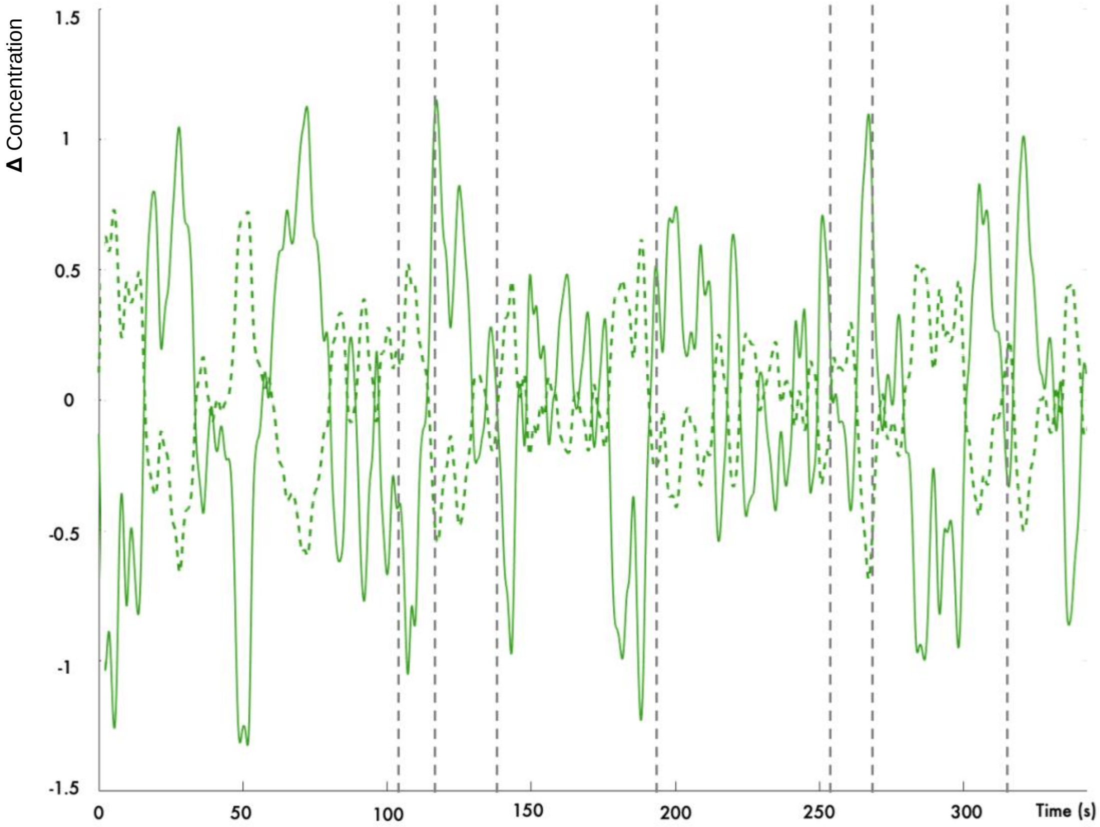


Figure 21 - Time course of HbO and HbR for channel 19 of participant 6.

Only the mental events are marked. HbO is shown with the solid, and HbR with the dashed green line.

In Figure 21, an example of preprocessed signal for one subject is reported. The preprocessing pipeline was effective in minimizing the noise components in the raw fNIRS signals, including slow trends, cardiac pulsation, and motion artifacts. In fact, increases in HbO and decreases in HbR can be observed, which are time-locked to the stimuli presentations (dashed lines).

6.9. Data analysis (fMRI)

Participants were scanned with a 3-Tesla Siemens Tim Trio MRI scanner with a 32-channel head coil and integrated parallel imaging techniques at the Centre for Cognitive Neuroimaging, University of Glasgow (CCNi), University of Glasgow. Functional images were acquired using an echo planar image (EPI) sequence [multi-band EPI, TR = 2000 ms, TE = 26 ms, 68 slices per volume, 2 mms isotropic voxels, no gap]. Structural images were acquired using a three-dimensional T1-weighted imaging sequence scan [1 mm isotropic resolution, TR = 2300 ms, TE = 30 ms, FA = 9°, field of view = 192 x 256 mm^2]; as well as a field map [3.28 x 3.28 x 3.3 mm voxels, TR = 488 ms, TE = 4.92 / 7.38 ms, FA = 60°, field of view = 192 x 192 mm^2].

6.9.1. Pre-processing (fMRI)

Results included in this manuscript come from preprocessing performed using fMRIPrep 1.5.2 (Esteban et al., 2019) [RRID:SCR_016216], which is based on Nipype 1.3.1 (K. Gorgolewski et al., 2011) [RRID:SCR_002502]. Some of the tasks and sessions that are referenced in this pipeline were part of a larger project (BOLDlight) and are not reported in this chapter. The following fMRI processing steps are also reported in a secondary analysis of these data (Hortensius et al., in preparation).

6.9.2. Anatomical data preprocessing

A total of two T1-weighted (T1w) images were found within the input BIDS dataset. All of them were corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008) [RRID:SCR_004757]. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, (Zhang et al., 2001)). A T1w-reference map was computed after registration of 2 T1w images (after INU-correction) using mri_robust_template (FreeSurfer 6.0.1, Fischl, 2012). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al., 2011, RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym].

6.9.3. Functional data preprocessing

For each of the 10 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A deformation field to correct for susceptibility distortions was estimated based on a field map that was co-registered to the BOLD reference, using a custom workflow of fMRIPrep derived from D. Greve’s epidewarp.fsl script and further improvements of HCP Pipelines (Glasser et al., 2013). Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9, Jenkinson & Smith, 2001) with the boundary-based registration (Greve & Fischl, 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al., 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox & Hyde, 1997) [RRID:SCR_005927]. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in [‘MNI152NLin2009cAsym’] space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (Behzadi et al., 2007) [CompCor]. Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components’ time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite, 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

6.10. First and second level analysis (fMRI)

The data was analysed in SPM12 (Wellcome Trust Centre for Neuroimaging, UCL, London) in MATLAB version 8.5 (The MathWorks Inc., Natick, USA, 2018b). First-level analyses were conducted by using a general linear model (GLM). The design matrix included the previously introduced mental and pain events (Richardson et al., 2018), as well as the following predictors of no interest (Hortensius et al, in preparation): framewise displacement, six head-motion parameters, and a subset of anatomical CompCor confounds (white matter and CSF decompositions). A standard hemodynamic response function was modelled, complying with the recommendations of Jacoby and colleagues (2016). A grey matter mask was used with a threshold of 0.8. For the mental > pain contrast images were smoothed using a 5mm smoothing kernel. For the second level (group level) analysis, a one-sample t-test was computed (p <0.001 uncorrected, k=10). The ROI analysis was done by extracting contrast values for the left and right TPJ using the MarsBaR toolbox (Brett et al., 2002) in MATLAB (version 8.5) (Mathworks, Natick, MA, USA, 2018b). A 9mm sphere was built using the same coordinates (LTPJ: x = -48, y = -62, z = 30 and RTPJ: x = 48, y = 60, z = 30 in MNI space) as Richardson and colleagues (2018).

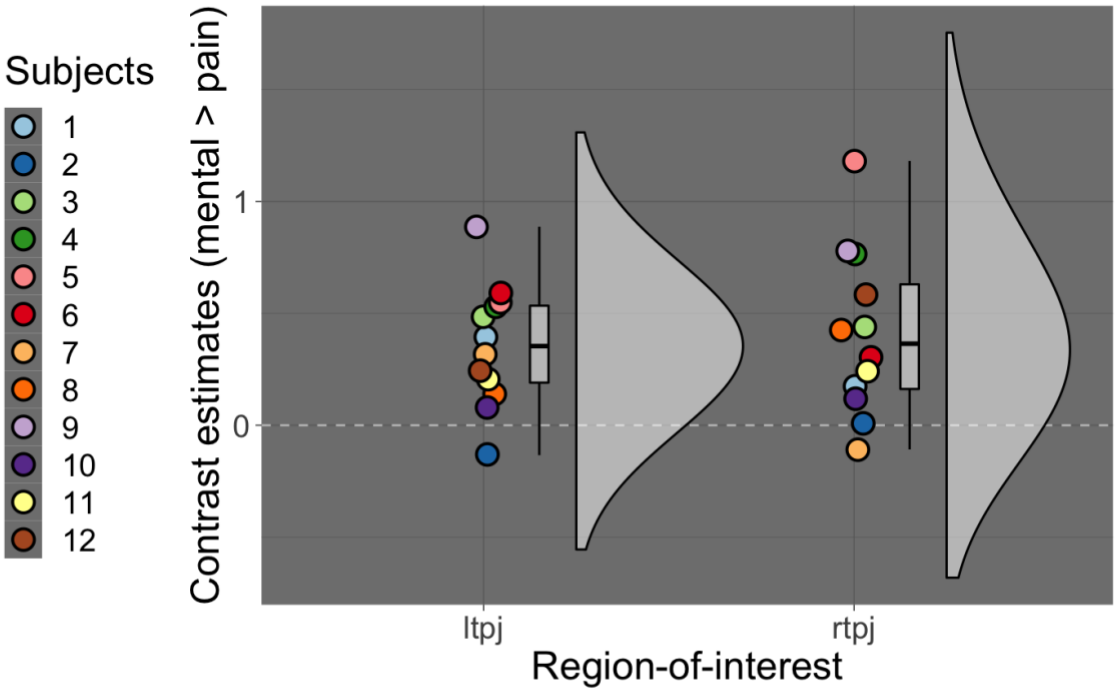


Figure 22 - Region of interest analysis (fMRI).

The region-of-interest analyses showed robust activation for mental events compared to pain events across the ToM network, specifically the left and right TPJ. Contrast estimates for Mental>Pain (left and right TPJ) are shown on the individual subject level. Colourful dots represent each participant's estimate. The dashed line shows the contrast level of 0.

6.11. Results

6.11.1. fMRI: Task validation

The localiser was successful, as it evoked robust activation of the ToM network for the mental > pain contrast, as shown by Figure 22 on the individual level for bilateral TPJ, and Figure 23 on the group level.

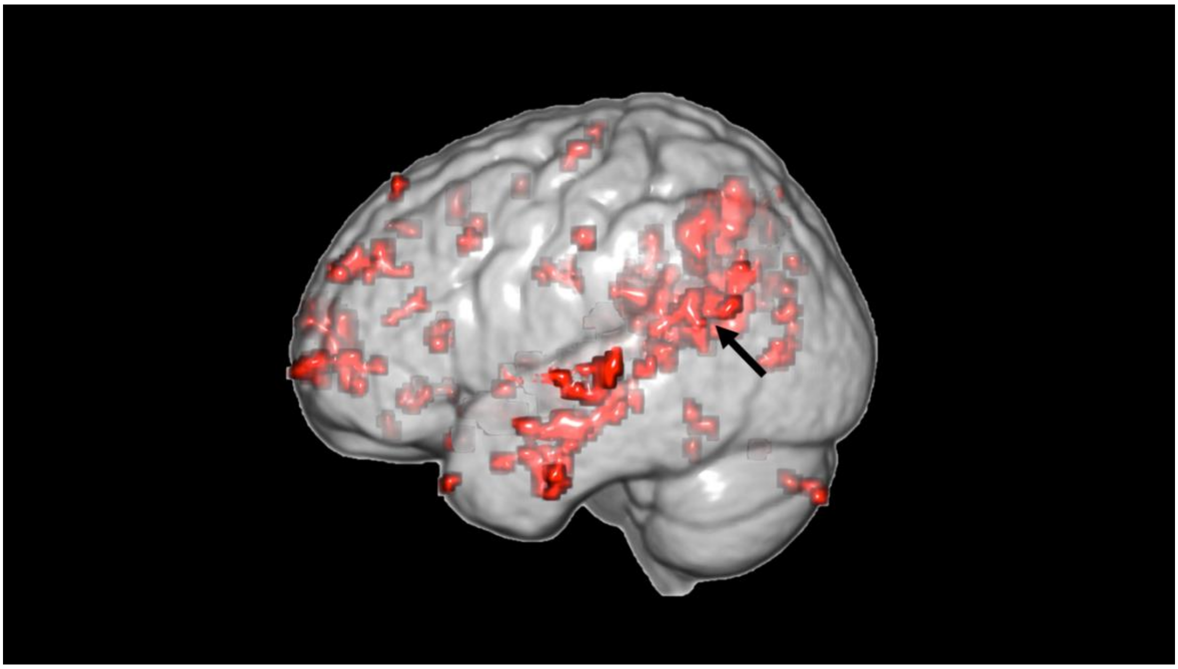


Figure 23 - Group map, left side (fMRI)

The group map shows robust engagement of the ToM network for the Mental > Pain contrast (in red) across all participants (middle temporal lobe, superior temporal sulcus, left TPJ (arrow), ventro- and dorsolateral medial prefrontal cortex).

6.11.2. fNIRS: Contrast effects

As illustrated in Figure 24, we report the uncorrected single-subject level results for the contrast of interest (Mental > Pain). Channels that were significant at the uncorrected level for *p* <.05 are indicated in green, channels that were not active are indicated in grey. These results show that there were significant changes in HbO only for a subset of participants: the array over the left side of the putative lTPJ shows activation for 7 out of 12 participants, for the putative rTPJ (as measured with the fNIRS system), 4 out of 12 participants show significant changes.

We speculated that due to the participants seeing the movie task at least twice (as fNIRS and fMRI was not measured concurrently, but subsequently), maybe an anticipatory shift was occurring. This has been reported in a recent study by Richardson and Saxe, who presented the same localiser to their participants twice (2020). To investigate this, we conducted an exploratory second analysis for those participants (#1-5, and #12), shifting the events of interest forward for two seconds (i.e. one TR, as reported by Richardson & Saxe, 2020). However, we observed the same response pattern, with the only difference of participant 3 showing no active channels on the left side, and for participant 5 on the right-side channel 6 was no longer significant.

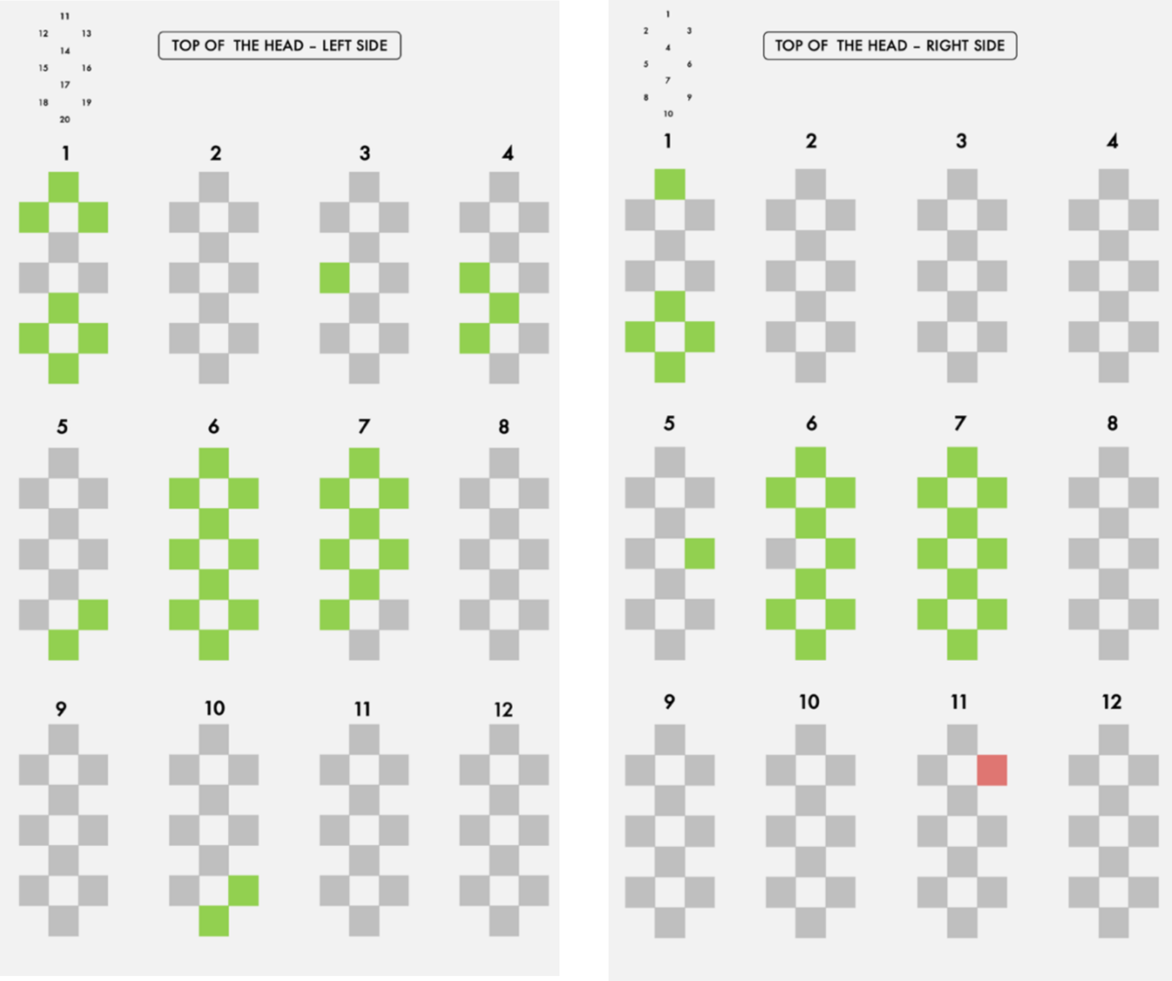


Figure 24 – Channels significant at the .05 significance level for the Mental>Pain contrast.

Channel 3 (subject 11) was excluded from the analysis. The original event timings were used here. More details of the analysis on shifted event timings can be found in section 6.11.2.

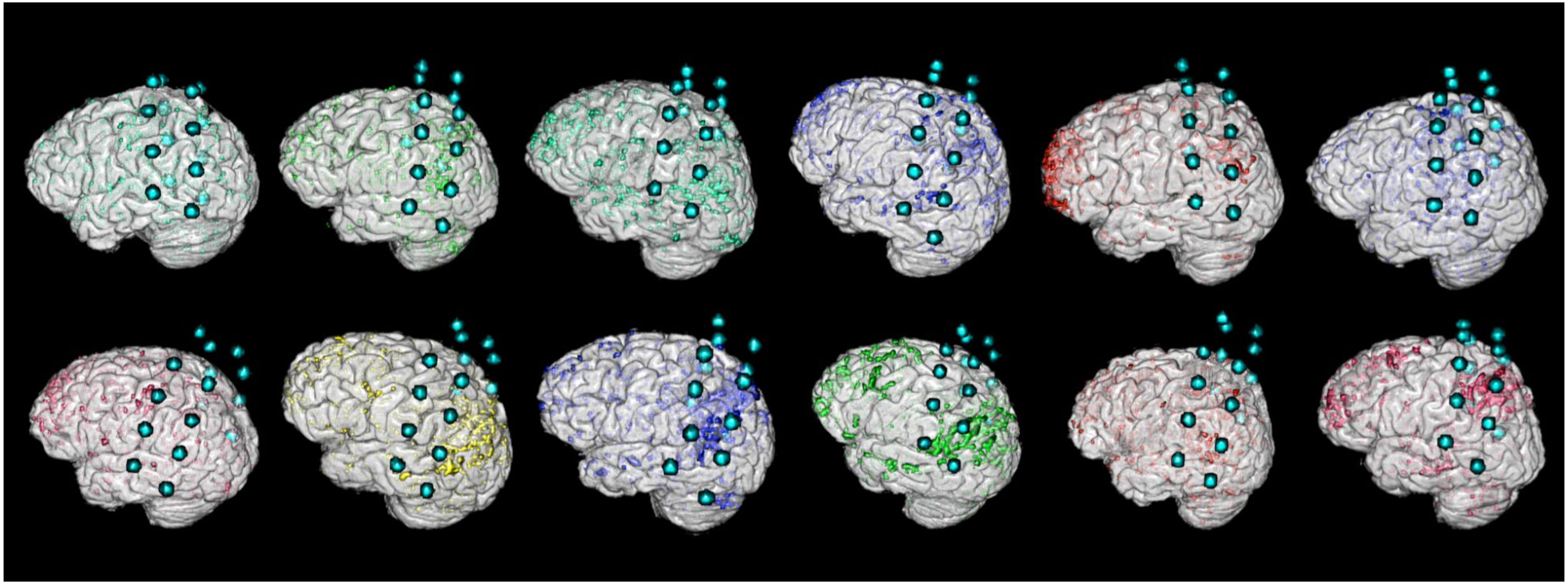


Figure 25 - Composite images consisting of each subject's anatomical image, functional activity and co-registered optodes in native space.

A closer look at Figure 24 reveals no consistent pattern of active channels: for some participants, only the area under the dorsal or ventral channels shows activation, for other participants (number 6 & 7), all channels are active for the contrast of interest. After checking whether an anticipatory shift had taken place, we turned to our overlap sensitivity analysis, to investigate how well the co-registered optode positions corresponded with the subject's functional activity in native space.

6.11.3. Overlap sensitivity (fNIRS/fMRI)

In Figure 25, we show the composite images resulting from the first level MRI analysis. Here we can see the optode positions that were extracted with the MarsBaR toolbox on top of individual participants' anatomical and functional images. For 10 out of 12 participants (>83%), we see a good overlap between the fMRI functional activity and the positions of the channels. However, inspecting individual subjects we also see large variability in the areas of covered cortex and arrangement of the optodes. A further important point, which becomes evident in this figure, is that some subjects show a much larger distance between the cortex and the optodes than others. This could explain some of the 'quiet' arrays we have shown in the previous Figure 24. Overall, Figure 25 shows large inter-subject variability of functional activation, anatomical structure and variation in optode placement and brain areas covered.

6.11.4. Detection rate (fNIRS/fMRI)

Comparing the detection rate on an individual subject level of the two modalities revealed that, as in the original study (Jacoby et al., 2016), the functional localiser was successful in eliciting activity in the expected functional ROIs (Figure 22). The left TPJ fROI was successfully identified for 11 out of 12 participants, and the right TPJ fROI was significantly active for 10 out of 12 participants. Hence, the detection rate of the fMRI functional localiser task was more successful compared to the fNIRS modality (7 out of 12 participants for the left-side probes, 4 out of 12 for the right-side probes).

6.12. Discussion

In this proof-of-concept experiment, we set out to answer two questions: (1) how well would optode placement, which was guided by participants' identified fiducial points, overlap with the anatomical and functional MRI images; and (2) how would the detection rates of fMRI and fNIRS compare? In addition, we sought to investigate how feasible adapting a new method for spatial registration from the EEG to the fNIRS modality would be. We answered these questions by taking a descriptive approach, visually inspecting the composite images consisting of co-registered optode positions, the subject's anatomy and functional activity. For most participants the optodes seemed to cover the brain activity recorded with fMRI, but we observed many interindividual differences relating to scalp-brain distance, location of the functional activity and inconsistent placement of the optodes. Comparing the results of the first-level analysis, we were able to replicate a high detection rate in the fMRI modality, showing that the localiser task elicited activity of the ToM network at the individual subject level. However, our fNIRS channel-wise analysis showed less consistent increases in HbO. A little over 50% of the participants had any active channels on the left side of the head, and even fewer showed any increases in HbO on the right side.

One possible reason for a muted response could have been that upon viewing the movie for a second time, when fNIRS was the second experimental block, an anticipatory shift occurred. We investigated this possibility by shifting the events 2 seconds earlier in time for those participants who had seen the movie for the second time. This resulted in a very similar pattern, if not even more reduced than the first analysis we conducted, thus suggesting this possibility is less likely.

Another plausible explanation for the variability of the fNIRS activation is the scalp-brain-distance between the probes and the cortex. For those subjects with larger scalp-brain-distance, we can be almost certain to have not recorded any signal at all. Indeed, a relatively high attrition rate seems to be commonly reported in the fNIRS literature. For example, Plichta and colleagues (2006) observe good results at the group-level, however a poor outcome on the single-subject level. Ferrari and Quaresima (2012) highlight that fNIRS activity can be reliably reproduced even over years, at the group level. To summarise, further group-level analyses are necessary to definitively answer how successful the localiser was at evoking reliable increases in HbO. We can, however, confirm previous findings that on the subject-level, for an event-related social task, consistent activity is not observed. Thinking further about the proportion of active channels reported in the literature, it is often the case that in arrays of 40 or more channels, 1 or a maximum of 3 channels are reported that also survive multiple comparisons correction (Hyde et al., 2015; Powell et al., 2018; Wijeakumar et al., 2017).

Overall, physiological confounds may explain the results we report here. For instance, speculating further on why we did not observe consistent HbO increases on the individual subject level, one could also imagine that the repeat presentation of the movie stimulus could have led to repetition suppression – instead of an anticipatory shift (Bhandari et al., 2020; Larson et al., 2013). In fMRI and fNIRS studies, repetition suppression is a common approach to investigate the engagement of the same brain network (for example the putative pain network) for different types of experimental conditions, such as humans or robots experiencing pleasure or pain (Cross, Riddoch, et al., 2019; Nordt et al., 2016). Thus, future analysis could follow-up on possible suppression of the brain signal due to repeat presentation of the movie stimulus. Further, we cannot exclude the possibility of extracerebral noise playing a role in obscuring the signal, as we were constrained by the design of the cap and thus were not able to include short-separation channels, which have been proposed as a strategy to better account for different types of signal that does not originate from the brain (Tachtsidis & Scholkmann, 2016). Task-related systematic activity relating to heart rate, blood pressure, breath and the response of the autonomous nervous system contribute and obfuscate the true brain signal. As a solution, Tachtsidis and Scholkmann (2016) have proposed the use of these short separation channels, which can be used to partition the influence of the extracerebral activity.

6.13. Alternative analytical approaches for fNIRS

Reflecting on the perhaps not sensitive enough channel-wise analysis approach, we also consider alternative methods for future analysis. Indeed, Pinti and colleagues (2020a) implemented additional steps when processing the fNIRS signal, which were not used in this exploratory study. These authors applied correlation-based signal improvement (CBSI), and further down-sampled the signal to minimise the impact of serial autocorrelations on the GLM (Pinti et al., 2019). Beyond the standard array-based analysis we report here, another option can be to use an ROI approach, i.e. to consider groups of channels within an ROI – which, in the absence of individual anatomical scans can be guided by the fOLD toolbox (Zimeo Morais et al., 2018) and which may be more robust to the pruning of bad channels (Pinti et al., 2018). Olmen (2018), who was interested in identifying relevant channels for their analysis of rTPJ activity, estimated cortical sensitivity using the Monte Carlo photon migration simulation algorithm with Atlasviewer (a separate functionality of the Homer2 toolbox). With this method, one channel of interest was identified and a time-window of 2 seconds around the peak of the HRF was analysed.

Very closely related to the multiple or single channel ROI approach, is the promising functional channel of interest (fCOI) method proposed by Powell and colleagues (Powell et al., 2018). In their study, which used a similar movie-stimulus (Baby Einstein, Walt Disney Productions), the authors tested a new analytical method by which channels of interest are identified in individual subjects, and the response is then tested in an independent set of data. Powell, Deen & Saxe (2018) argue that if the contrast of interest is sufficiently specific (in their case, video clips containing faces versus scenes, in our case mentalising vs. pain events), instead of treating channels with the same array positions as equivalent, one could be guided by the functional response profile on the subject level to identify the channels of interest in the left out data. Comparing the array-based and the fCOI approach (for the HbO results), the authors reported that no channel survived multiple comparisons correction in the array-based approach for the adult sample, and only one channel survived in their infant sample. However, with the more sensitive fCOI analysis, the authors were able to show that when the anterior portion of the array was analysed, responses were significantly higher for face compared to scene trials for both adults and infants, confirming the finding that very young children under the age of 1 preferentially respond to faces relative to scenes. Powell and colleagues (2018) criticise the approach implemented in the current study, where overlap of the channels with underlying subject specific functional regions is checked, as the size and location of these functional profiles can be highly variable (something we have observed also in this study). Thus, accurately matching channels to specific functional regions is insufficient. However, an important point to consider is power, as both in our study and in the studies by Powell and colleagues (2018) subject numbers were small (12, 20, 16, respectively) and thus, as they have also argued, the possibility cannot be excluded that with a larger sample size, the inherent noisiness in the array-based approach could be overcome. Finally, the authors remark that limitations of the fCOI approach could be partially addressed by still mapping optode locations to anatomical images, where possible.

6.14. Spatial registration methods

On balance, spatial registration remains an important point when implementing fNIRS studies for HRI. Evaluating the feasibility of the photogrammetry method for fNIRS, we conclude that for the majority of participants, with two runs of picture-taking, we obtained excellent quality head models. Issues we faced were mainly related to the black-and-white design of the cap, which we addressed by adding colourful stickers to the cap and then manually selecting the probe locations in janus3D. This of course adds time to the processing procedure: researchers can budget about 20 minutes for the photography, and between 1 and 2 hours for the construction of the head models in Metashape (depending on number and quality of the pictures), as well as an additional hour rotating and aligning the head models with the anatomical scans in janus3D. This would be the most salient disadvantage of using this method compared to faster methods, like for example video-based construction of head models. When we were piloting the video-based method, the resulting head models did not look like an accurate representation of reality, so this approach was abandoned early on. Since 2019 however, technical advances have resulted in alternative spatial registration methods, that might be more fast-paced and especially useful for infant study participants.

Jaffe-Dax and colleagues used a GoPro with the slow-motion feature and 3D surface reconstruction software (Structure from Motion, Visual SfM) to build the head models. These authors added colourful stickers on the fiducial points and covered the fNIRS cap with a pattern of blue and pink colourful cut-outs. This more time-efficient method showed good correspondence between the optode positions obtained from the video-based source reconstruction and the 3D magnetic digitizer method. This method has the additional advantage of subjects being able to move freely *while* the video is taken.

Another technical innovation is the use of structured-light 3D scanners for the estimation of probe locations (Homölle & Oostenveld, 2019). These low-cost scanners (which are for example used in Kinect cameras for Xbox), also allow relatively fast (the authors estimate 2 minutes, compared to the 7 minutes needed with the Polhemus digitizer) mapping of probe locations. The authors also highlight that they tested the approach on 50 subjects, compared to the single-subject (or single replica head) that has been commonly observed in the spatial registration literature. Overall, the authors found that the structured light scanning method showed good overlap with the Polhemus-obtained positions, however that this strategy was also not flawless. In one out of 50 participants the transmission of the electrode positions from the iPad to the computer failed, and in one out of 50 participants the Polhemus locations were not on the scalp. Homölle and Oostenveld (2019) recommend the 3D structured light scanning also for the recording of the position of NIRS optodes, as in their study it yielded comparable results to the positions obtained with Polhemus.

Finally, another promising method could be to use virtual registration, based on simulations (Tsuzuki & Dan, 2014), or a further alternative (in the absence of structural MRI scans) could be to spatially register probe positions to age-matched templates (like the MNI template brain) avoiding the need for MRI scanning (Bulgarelli et al., 2018).

6.15. Limitations

Several limitations have to be acknowledged. The probe positions in our set up were obtained via photogrammetry, before the optodes were attached to the fNIRS cap. Experimenters went to great lengths to avoid movement of the cap when attaching the optodes, however, especially in those cases where the initial probe check flagged problematic sources or detectors, the removal and reattachment of the probes could have led to a small margin of error in the final co-registered composite images. This small registration error can be assumed to lie between 1 - 10mm based on the findings of Clausner and colleagues (2017), and Hu and colleagues (2020). Hence, the putative registration error is smaller than the fNIRS spatial resolution (Pinti, Tachtsidis, et al., 2020b).

Furthermore, our placement of the array was guided by standards in the previous fNIRS literature, however, a better strategy would be to obtain probabilistic brain regions and their corresponding optode locations by using the new open-source fOLD toolbox (Zimeo Morais et al., 2018). With the help of the toolbox, the initial selection of the array placement could be better guided by brain regions of interest, as identified by the simulated photon transport method this toolbox applies. Different parcellation methods can be used, but the results of the toolbox are currently still restricted to the 10-10 and 10-5 international cap systems.

Another important limitation in this study is the difficulty we encountered with the digitisation of participants' landmarks, which prevented the spatial normalisation step to MNI space and the planned mapping of all participants' optode locations, to inspect the variability of the probe placement quantitatively. Common approaches in the literature are either adding colourful stickers or felt-tip marker points on the subjects' fiducial points, to later extract this important information in the spatial co-registration step (Homölle & Oostenveld, 2019; Jaffe-Dax et al., 2019). This information is missing for many participants in the current study, so we could not take this part of the analysis further.

6.16. Planned analyses for BOLDlight

Future analyses for BOLDlight will include group-level analysis of the fNIRS functional localiser task to compare how the channel-level detection rate compares to the overall fNIRS literature. Further, the HbR response will also be taken into consideration, as well as the multiple comparisons corrected results. Moreover, as we have a second dataset of the same task available for the second session that participants underwent for the re-test, we will investigate the feasibility of the fCOI approach outlined by Powell and colleagues (2018). In addition, we will aim to probe the spatial specificity of fNIRS utilizing a finger and foot tapping task, as well as a separate functional superior temporal sulcus localiser task (Isik et al., 2017).

It will be interesting to compare the results from these block-design tasks to the currently presented event-related functional localiser. Taking into consideration also the second scanning sessions for both modalities, we will establish and compare test re-test reliability for both fNIRS and fMRI, as recently concerns have been raised on the reliability of the fMRI modality (Elliott et al., 2020). Future experiments following on from BOLDlight will adhere to the initially proposed stepwise procedure of moving “cognitive neuroscience [...] from lab to life” (Henschel, Hortensius, et al., 2020b; Pinti et al., 2018, p. 369). We plan to conduct a direct replication of Walbrin and colleagues’ experiments (2018) using social versus non-social animations of interacting geometric shapes, and then move to a more ecologically valid stimulus set of robots and humans (Brough, Henschel, Rabagliati, Harris, Cross, & Branigan, 2020: Scotbots database, in preparation).

6.17. Acknowledgements

We would like to thank Paola Pinti and Katharina Stute for their consultation on fNIRS data collection, pre-processing and analysis. Further, we would like to express our gratitude towards Tommy Clausner and Daniel Livingstone for their valuable support with implementing janus3D for optode spatial registration. We gratefully acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement number 677270 to E.S.C.), the Leverhulme Trust (PLP-2018-152 to E.S.C), and the Bial Foundation (to R.H.).

6.18. Author contributions

Conceptualization: Anna Henschel, Ruud Hortensius and Emily S. Cross.
Data Curation: Anna Henschel and Ruud Hortensius.
Formal Analysis: Anna Henschel, Ruud Hortensius, Michaela Kent, Paola Pinti, Katharina Stute and Luca M. Leisten.
Funding Acquisition: Ruud Hortensius and Emily S. Cross.
Investigation: Anna Henschel, Ruud Hortensius, Luca M. Leisten, Hanna Seelemeyer.
Methodology: Anna Henschel, Ruud Hortensius, Michaela Kent, Luca M. Leisten and Hanna Seelemeyer.
Project Administration: Anna Henschel, Ruud Hortensius and Emily S. Cross.
Resources: Anna Henschel, Ruud Hortensius and Emily S. Cross.
Software: Anna Henschel, Ruud Hortensius and Luca M. Leisten.
Supervision: Anna Henschel, Ruud Hortensius and Emily S. Cross.
Validation: Anna Henschel and Ruud Hortensius.
Visualization: Anna Henschel, Ruud Hortensius, Michaela Kent and Luca M. Leisten.
Writing - Original Draft Preparation: Anna Henschel and Michaela Kent.
Writing - Review & Editing: Anna Henschel, Ruud Hortensius and Emily S. Cross.

CRediT taxonomy by Holcombe and colleagues (2020).

6.19. Citation Diversity Statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field (Dworkin et al., 2020). Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al., 2020). By this measure (and excluding self-citations to the first and last authors of our current paper), **our references contain 16.22% woman(first)/woman(last), 12.16% man/woman, 14.03% woman/man, and 57.59% man/man**. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, non-binary, or transgender people. We look forward to future work that could help us to better understand how to support equitable practices in science.

Adapted from Dworkin and colleagues (2020).

Chapter 7 General discussion

The primary aim of this thesis was to develop robust behavioural and neural methods for the investigation of interactions with humanoid robots. Based on insights stemming from psychology and neuroscience, I adapted paradigms for human-robot interaction, with the ultimate goal of conducting more reliable and ecologically valid experiments probing social motivation towards robots. Across the chapters outlined in this thesis, I uncovered important issues in adapting these paradigms for HRI research. Below, I summarise the main findings and contributions, and critically reflect on the limitations of the present work. In the sections following the general discussion of the results, I summarise methodological implications and highlight recent theoretical considerations. Here, I especially highlight the impact of the replication crisis and a move towards incorporating open science methods across disciplinary boundaries.

7. Summary, Contributions & Limitations

In **Chapter 1**, I presented a general overview of the field of social robotics and its historical development. While reviewing popular definitions of social robots, it became clear that the terminology for referring to these machines is far from clear-cut. However, most authors describe a social robot as an embodied agent that is able to communicate and collaborate with humans in a socially engaging way. My review of these definitions also showed that many researchers framed social robots in future-oriented terms, with the currently available machines perhaps not quite meeting the ideal vision of a truly sociable robot. I also reviewed mainstream robotic platforms and highlighted their fields of application, which revealed that social robots are expected to be deployed in care scenarios, functioning as companions for the elderly and as social skills teachers for children with ASC. The introduced robots demonstrate the heterogenous morphology of currently available commercial and bespoke research systems. In the final part of the chapter, I discussed the social robot paradox, which describes a mismatch of expectations and reality when it comes to the limited success of integrating especially *humanoid* robots into the human social ecosystem.

7.1. Humanoid robots to advance our understanding of social cognition

Despite this limitation, humanoid robots are useful research tools, a point I highlighted in **Chapter 2**. Using robots in cognitive neuroscience and psychology experiments allows researchers to explore how these agents, which occupy a liminal space in terms of their socialness, appeal to cognitive systems that developed over the evolution of successful human-human interaction.

Humanoid robots also help address a major challenge social neuroscience is facing at the moment. The investigation of social perception and behaviour has been a primary aim, but experiments have been historically limited by presenting mainly screen-based social scenarios (Schilbach, 2012). Bringing humanoid robots into the lab allows researchers to take advantage of excellent experimental control (to a fine-grained degree), and at the same time achieves better ecological validity with embodied, co-located social agents.

Finally, an important advantage of using social robots in an experimental context is that it forces a close examination of fundamental questions, thereby contributing to an overall advanced understanding of human social cognition (Bolis & Schilbach, 2018). In **Chapter 2** I also discussed the tensions that emerge at the heart of these interdisciplinary efforts. Given these tensions, some researchers have emphasized that a strong theoretical foundation is crucial when adapting paradigms for HRI (Eyssel, 2017). After reviewing popular theories in HRI, I gave an overview of the framework that underlies this thesis: The Social Motivation Theory of Autism (Chevallier et al., 2012). This theory set the stage for the following empirical pieces of work, which address two fundamental pillars of the framework: the reward value of social interaction and attentional capture by salient social cues.

Chapter 3 was motivated by an observation described in the previous two chapters: although many social robots are already deployed in contexts with vulnerable users, they often fail to deliver on their promise to adequately and autonomously respond in social situations. Other researchers in the field have similarly observed that features and behaviours are often implemented in social robots before they are evaluated as being beneficial or crucial for facilitating engaging, long-term interactions (Vallverdú & Trovato, 2016).

In this study, we tested how movement synchrony with a robot would affect its likeability and participants' motivation to spend more time with it in a free, conversational interaction. However, after reviewing meta-analyses on the effects of interpersonal synchrony on rapport in human-human interaction, we identified a problematic confound highlighted by Rennung and Göritz (2016) that could obscure the true experimental effect of synchrony on factors such as likeability or perceived social cohesion. Many studies failed to ensure experimenter blinding, which is particularly problematic as it has been recently established that experimenter beliefs influence participant behaviour in experimental contexts (Gilder & Heerey, 2018).

One major contribution of this work is that we ensured a double-blind experimental procedure, where neither participants, nor the experimenter, were aware of the experimental condition. We found that participants who had synchronised with the Pepper robot in a drawing task did not rate the robot as more likeable, intelligent or more human-like compared to the group of participants that did not synchronise with the robot on this task. Further, we did not see that participants in the synchrony group chose to ask the robot more questions in a semi-structured conversational interaction scenario that was presented as an optional part of the experiment. Across the two groups, participants were equally motivated to engage in an interaction with Pepper, as the number of questions the participants asked did not differ statistically based on their group assignment. An interesting finding that we observed in this experiment was that there was a discrepancy between objectively manipulated and subjectively perceived synchrony. A larger proportion of participants in the asynchrony group stated that they perceived to have been in sync with the robot - which highlights the need for including careful manipulation checks in these types of experiments.

Another contribution of this study relates to the lessons we learned using humanoid robots in experimental contexts. Through extensive piloting prior to data collection, we established that the autonomous mode of the robot was not able to successfully sustain interactions with different participants, especially under changing light conditions and when participants spoke with varying accents.

When we piloted the free interaction scenario, we (anecdotally) observed that participants quickly felt rejected by the robot when it did not immediately orient towards them, when spoken to. This moved us to take advantage of remote controlling the robot via a Wizard-of-Oz (WoZ) procedure, where the gaze of the robot was always directed at the correct angle towards the participant, and it responded to the questions without any larger delays, as the answers were triggered by the experimenter behind a room divider. An important limitation of this chapter is that we did not outline in detail the scope of the perceptual and cognitive WoZ control according to the guidelines proposed by Riek (2012). Future studies should seek to report details of wizard training (e.g. when experimenters of this study practiced the control of the verbal utterances of the robots) and any potential experimenter error (e.g. if there was any delay in eliciting the answers of the robot) in more detail.

Another observation, which is interesting in the light of how the Pepper robot was designed (see **Chapter 1**), was that in order to successfully set up the drawing task with the participant and the robot side-by-side, we had to take into account the robot sensors, which would lead to freezing its motions if the participant, the screen or the table were positioned too close to one another.

Some additional challenges have to be acknowledged and I will briefly discuss these in the following. We did not pre-register our analytical plans and the statistical analysis of the null result was limited to a Frequentist approach. To investigate whether these data provide evidence for the absence of an effect of the synchrony manipulation, an additional Bayesian analysis would have been desirable.

A further limitation, which was addressed by reviewers of the manuscript (Appendix B), relates to the fact that we did not assess the extent to which participants attributed mind or intentionality to the robot. While the robot introduced itself as a member of the research department, the influence of the synchrony manipulation could have been obscured by the participants not finding the robot believable as an intentional agent. Anecdotally, one participant found the robot stating that it did not have a birthdate so devastating, that they offered to bake it a cake. Unfortunately, these spontaneous responses were not recorded, and point to a need of implementing both quantitative and qualitative measures of evaluating interactions in HRI to capture a complete picture of perceived mind and intentionality (Riddoch & Cross, 2020).

Another limitation of the study reported in **Chapter 3** was the fact that no positive control in the form of a human interaction partner was included. In the light of the findings described in the subsequent chapters, I will make the case that in addition to including positive controls, it will be important to conduct direct replication studies of original effects – and where possible, to go one step further by using the Registered Replication Report format (Simons et al., 2014). These ideas are outlined in more detail in the following sections of this chapter.

To evaluate the robot, we used self-report measures developed and validated for embodied HRI scenarios (Bartneck et al., 2009; Nomura et al., 2005). However, this study uncovered important limitations in using these questionnaires: indeed, we detected that subscales of the popular Godspeed Questionnaire Series contained duplicate items, which led to the inclusion of only 3 of the 5 subscales (Weiss & Bartneck, 2015). The limitations of the Godspeed questionnaire have been highlighted by other researchers as well (Shen et al., 2015), and in conjunction with the null result, a distinct lack of robust behavioural measures to quantify social motivation towards robots became apparent (Chevallier et al., 2016). We originally attributed the lack of an observed effect to our interaction measure as not sensitive enough to capture the subtle experimental manipulation. This led to the empirical study described in **Chapter 4.**

7.2. Towards robust behavioural measures of social motivation

In **Chapter 4**, we investigated the social relevance of humanoid robot faces by means of conceptually extending the eye-contact effect (Conty, Gimmig, et al., 2010a). Conty and colleagues (2010a) found that when presenting socially salient cues as distractors during a demanding cognitive task, the cues would impair the performance of participants by slowing down reaction times over and above of what is normally expected for the Stroop interference effect. This task has been proposed as a behavioural “proxy for social motivation” (Chevallier et al., 2013, p. 1694).

In Experiment 1, we found the expected Stroop interference effect, indicating that despite the modifications we made in this conceptual extension, the task worked as expected.

The main contribution of this chapter relates to the extension of the eye contact effect (Conty, Gimmig, et al., 2010a; Senju & Johnson, 2009) with a newly collected stimulus set, that included well-controlled greyscale images of human faces (DeBruine & Jones, 2017; Langner et al., 2010), humanoid robot faces, object faces and non-social control images of flowers. Chevallier and colleagues (2013) anticipated that the effect would extend to other cues, such as whole faces, but we cannot support this idea with the data collected in our experiments.

A small, distractor-dependent interaction emerged in Experiment 1, yet this effect disappeared in Experiment 2, which was adequately powered to detect the effect size of interest. We found no differences across the four different stimulus categories of capturing attention due to their more or less social nature. Thus, our main conclusion across these two experiments was that despite previous literature on social attentional capture, we were not able to adapt this effect for HRI. Indeed, we failed to show that the most salient social distractor, the human faces, would robustly captured participants’ attention. Our findings initially appeared at odds with the established literature around this effect. But even so, recent studies by Pereira and colleagues (2019, 2020), which carefully controlled for known confounds, also failed to show a social attentional capture by human facial cues.

Some limitations to this empirical work must also be acknowledged. While we pre-registered the processing of the reaction time data, we did not take into account the garden of forking paths inherent in this procedure. One example of this can be seen in the multitude of available processing methods in the R-package ‘trimr’ (Grange, 2015). Our method of pre-processing the reaction times involved using a standard deviation reaction time trimming criterion, which was criticised by a peer reviewer (Appendix D). To investigate whether a conservative exclusion of experimental trials played a role, we used an alternative, participant-sensitive standard deviation criterion. Still, this resulted in similarly shaped reaction time distributions (Appendix D).

A further challenge was determining adequate limits for the region of practical equivalence with zero approach (ROPE) used in the exploratory Bayesian modelling analysis. After initially using an automatic procedure via a function built into the 'BayesFactor' package (Morey et al., 2018), the resulting ROPE range was too large, so that we determined it based on half of what we consider a small effect (Kruschke, 2018). Contrary to our expectations, the analysis was inconclusive. We were not able to find strong support for the null hypothesis, as the posterior samples only partially overlapped with the ROPE. We can only speculate about the nature of this inconclusive result; however, one underlying reason might have been the strategy we chose to determine the bounds of the ROPE.

Together with the recent work by Pereira and colleagues demonstrating the influence of confounds in measuring social attentional capture (2019, 2020), we can conclude that the task is not suitable for measuring social motivation in human-robot interaction studies. Whereas covert measures of social attention may not be a fruitful avenue for HRI researchers, overt measures, such as eye-tracking, may yield more informative findings (Hayward et al., 2017).

7.3. Mobile brain imaging to facilitate embodied experiments with humanoid robots

In **Chapter 5**, my co-authors and I provided an intermediate reflection and opinion on the current state-of-the-art neuroscience tools in embodied human-robot interaction studies. Overall, we foresee that the field will be driven forward by more transparent, embodied and mobile neuroscience. Here we especially highlighted the 'promises and pitfalls' of fNIRS, which affords the advantage of allowing brain imaging during embodied and interactive encounters with social robots. Recent advances in the development of portable and lightweight fNIRS systems can support more ecologically valid experimental interaction paradigms.

Finally, my co-authors and I argue that it will become important in the coming years to move beyond a predominantly anthropocentric approach and consider other comparison categories (for example pets or objects), as well as brain regions beyond popular ‘hub regions’ of social cognition. I will reflect on this point in the final sections of this thesis discussion.

In this chapter, we also highlighted the repercussions of the replication crisis, and the move towards open science practices in psychology and neuroscience. Especially when implementing fNIRS in embodied human-robot encounters it is important to follow a stepwise approach, starting with a replication of effects. We argued that it is in the interest of various stakeholders that research builds on strong foundations, using rigorous and robust methods.

Chapter 6 implemented the proposed stepwise approach of the previous chapter: here we sought to validate a novel mobile fNIRS system, by comparing this brain imaging modality to the current gold-standard in social neuroscience: fMRI. A second contribution of this chapter was the adaptation of a new, photogrammetry-based method to co-registering the positions of the fNIRS probes. The main focus of this chapter was to transparently describe and highlight challenges encountered when using this brain imaging method, especially in comparison with the results yielded by fMRI.

We investigated two main questions: the first being the overlap sensitivity between the placement of the optodes and the functional regions of interest – as determined by fMRI. Secondly, we were interested in the detection rate of each modality at the single-subject level. After extensive piloting we constructed a reliable photogrammetry pipeline that resulted in 11 out of 12 excellent quality head models. We provide a detailed description of every step in the openly available manual (Henschel, Kent, et al., 2020c). This new method to digitise optode positions was only successful after adding colourful markers to the fNIRS cap. We found that the automatic identification algorithm of the janus3D co-registration toolbox (Clausner et al., 2017) was not reliable for the fNIRS optode holders (as opposed to the EEG electrode holders it was designed to detect).

Consequently, the experimenters had to choose a more time-consuming route of manual tagging the optode locations in janus3D.

While we found that the position of the optodes overlapped well with participants' functional MRI activity, this method, as highlighted by Powell and Saxe (2018), is not without issues. It is challenging to support the success or failure of probe location overlap based on single-subject functional activity, as anatomical and functional locations were highly variable. The detection rate on the single subject level was better for fMRI, as compared to fNIRS. Indeed, by showing composite images of the optodes mapped to subject anatomical and functional MRI images, we identified that strong inter-subject variability in terms of scalp-brain distance was one of the contributing factors. However, we also investigated the possibility that an anticipatory shift had taken place, as a recent paper by Richardson and colleagues had shown that repeat viewing of this particular localiser movie led to a predictive response of the brain (2020). Shifting the events forward in time did not lead to more observed fNIRS activity; on the contrary, the activity was even more subdued.

Overall, we can conclude that on the single-subject level, the signal recorded with fNIRS showed a lower signal to noise ratio than the fMRI signal, which replicated reliable activation of the left and right TPJ almost 100% of the time across all subjects. This point is important, as future studies will seek to implement the fNIRS system in mobile interaction paradigms with robots, and thus through more subject movement, noise levels might further increase. In the current task, subjects sat still and passively observed a movie, which still only resulted in a success rate of elicited fNRIS activity in just over half of the participants. This will be a crucial factor when planning subject recruitment numbers, and when calculating power for future studies.

7.4. What does the replication crisis mean for the future of HRI research?

Most empirical findings reported in this thesis are null results, which are inherently challenging to interpret due to the complex factors that could be at play contributing to the failure of conceptually extending an effect (Shrout &

Rodgers, 2018; Simmons et al., 2017). Publishing null results has become more acceptable in the research community in recent years, as more and more scientists grappled with replicating well-known effects. This led to the replication crisis in psychology and neuroscience (Aarts et al., 2015; Schimmack, 2020; Shrout & Rodgers, 2018). As is perhaps evident through examples presented in this thesis, the ripples of these crises have not been limited to psychology and neuroscience. They are reaching HRI as well, exacerbating the interdisciplinary tensions discussed in **Chapter 2** (Irfan et al., 2018).

Some researchers note that trust in the ‘social sciences’ has been shaken as a result of the crisis (Irfan et al., 2018), others have a more optimistic outlook: indeed, as Shrout and Rodgers (2018) argue, the sense of urgency invoked through the imagery of a crisis has sparked fundamental and sustained positive change. The authors highlight how, as a result of this crisis, some of the open science movement’s most valuable tools have been created: the Open Science Foundation (Nosek, 2013) and a global shift in research conventions, such as the increased adoption of pre-registering hypotheses, design and analysis plans (2018). As Schimmack (2020) writes in his perspective on replication ‘failures’ (where a recent Nature Human Behaviour editorial argues strongly for the fact that “replications do not fail”; Kousta, 2020, p. 559), the crises in psychology and neuroscience might serve as an important warning message to other disciplines, who have not yet felt the severe repercussions riding on their back.

7.4.1. Adoption of open science methods among the HRI community

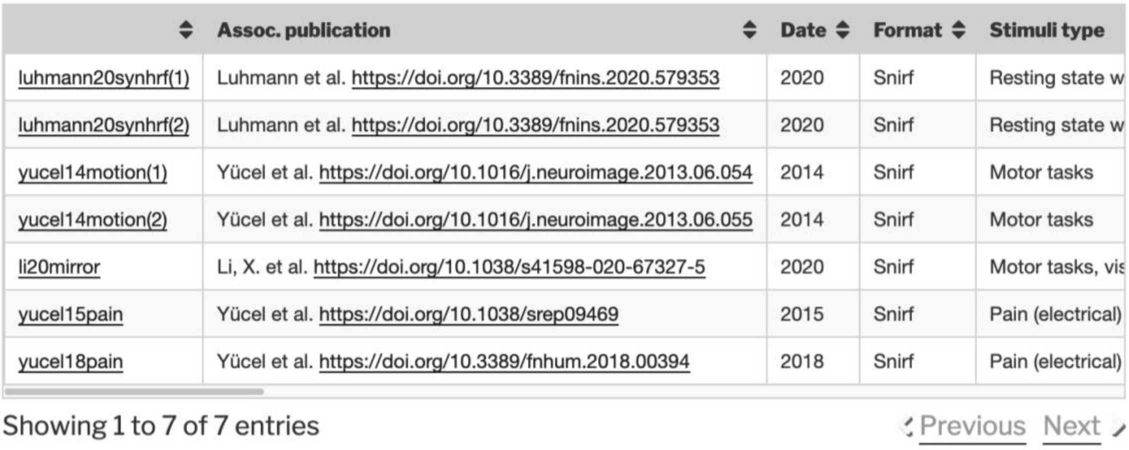
Researchers in HRI have acknowledged this crisis and have started to adopt some of the practices that scientists in related fields have been lobbying for over the past nine years (Baxter et al., 2016; Belpaeme, 2020; Irfan et al., 2018; Schimmack, 2020; Strait et al., 2020). Baxter and colleagues (2016), who analysed three years of proceedings stemming from the field’s most important meeting, the ACM-HRI conference, note that they observe a clear trend towards the adoption of open science practices: for example, journals such as PLOSOne now require datasets to be shared openly. Taking even more positive steps in this direction, in 2020, ACM-HRI for the first time *explicitly invited* replication studies in a dedicated conference track, with five proceedings replicating their studies across robotic platforms and data collection sites (Kubota et al., 2020;

James Li et al., 2020; A. Pereira et al., 2020; Sandygulova et al., 2020; Strait et al., 2020).

Strait and colleagues (2020) conducted a conceptual extension and three-site replication of the robot-adapted Joint Simon Effect (JSE), which examines how people represent the actions of a robotic co-actor (Stenzel et al., 2012). In their collaborative replication effort, the authors found that the expected JSE replicated. The authors noted that a key component to the success of this effort was that all three, international replication sites had access to the same platform: the Nao robot (Strait et al., 2020). This speaks to an observation described in **Chapter 2**: commercially available humanoid robots offer many advantages to HRI researchers, despite their limited social abilities 'in the wild'. Strait and colleagues (2020) conclude with the reflection that replicating effects will become ever more important in the HRI community, especially given the fact that attempts of running source code from the 2017 IEEE International Conference on Robotics and Automation (ICRA) proceedings were crowned with success in only 2% of the cases (Cervera, 2019).

7.4.2. Adoption of open science methods in the fNIRS community

However, HRI is not the only research community to be weighing up the next steps in a move toward more open and reproducible research practices: the relatively young field surrounding fNIRS research has been relatively slow to adopt open science conventions already embraced in other neuroimaging communities, for instance among fMRI researchers (Bratt, 2017). Currently, there are not many open fNIRS data sets available, and perhaps as a consequence, meta-analytical investigation of the robustness of social and cognitive tasks measured with fNIRS are scarce (Bendahan et al., 2019). This is mainly due to the heterogeneity of fNIRS devices and data processing methods used – for example Bendahan and colleagues (2019) sought to conduct a meta-analysis on the connection between cerebral oxygenation and the presence of delirium in patients. The meta-analysis could not be completed due to a large variability in fNIRS systems and pre-processing procedures. Increased use of standardized localiser tasks in fNIRS could lead to a better comparability across studies, and perhaps an increased motivation to share data. As of October 2020, there are only 7 datasets shared on OpenFNIRS (Figure 26, https://openfnirs.org/data/).



	Assoc. publication	Date	Format	Stimuli type
luhmann20synhrf(1)	Luhmann et al. https://doi.org/10.3389/fnins.2020.579353	2020	Snirf	Resting state w
luhmann20synhrf(2)	Luhmann et al. https://doi.org/10.3389/fnins.2020.579353	2020	Snirf	Resting state w
yucel14motion(1)	Yücel et al. https://doi.org/10.1016/j.neuroimage.2013.06.054	2014	Snirf	Motor tasks
yucel14motion(2)	Yücel et al. https://doi.org/10.1016/j.neuroimage.2013.06.055	2014	Snirf	Motor tasks
li20mirror	Li, X. et al. https://doi.org/10.1038/s41598-020-67327-5	2020	Snirf	Motor tasks, vis
yucel15pain	Yücel et al. https://doi.org/10.1038/srep09469	2015	Snirf	Pain (electrical)
yucel18pain	Yücel et al. https://doi.org/10.3389/fnhum.2018.00394	2018	Snirf	Pain (electrical)

Showing 1 to 7 of 7 entries

Figure 26 - The Openfnirs database (https://openfnirs.org/data/).

Another recent development in the fNIRS community may prompt an increased uptake of open science practices, as the sNIRF format has been introduced (https://github.com/fNIRS/snirf). Encouragingly, a brain imaging data structure (BIDS) extension proposal for NIRS has been raised (BIDS Extension Proposal, BEP030), which implements the file-naming and file-structure convention initially developed for fMRI (and which was used in **Chapter 6** to run the standardized pre-processing pipeline fMRIprep). The BIDS format, which is essentially a standardized brain data management plan, has the potential to greatly enhance reproducibility and open science efforts (Gorgolewski et al., 2016). The extension of BIDS for NIRS will incorporate the already developed sNIRF raw data format, which can be converted from most proprietary manufacturer file formats (e.g. in **Chapter 6** we initially translated the Shimadzu files to the nirs format).

Bratt (2017) laments a lack of open data repositories for fNIRS (in studies on emotion – yet, as Figure 26 shows, this is a global problem), in particular relating to the 'curse of dimensionality' known in the machine learning field. Researchers who may want to use machine learning procedures to analyse their data, are faced with small samples in fNIRS studies. Using machine learning classifiers on highly dimensional data, such as the fNIRS signal, is challenging as these data have many features with relatively few observations (Jie Li et al., 2016). As a result, models are 'overfit' and are not generalisable beyond individual datasets (Bratt, 2017). One way to address this issue (and in the future implement machine-learning methods for fNIRS) may be the aggregation of large, open datasets. In addition, Bratt (2017) argues that while fNIRS and FMRI are facing validity issues, increased data sharing, including text-based metainformation (see Neurosynth, https://neurosynth.org/), also has the potential to facilitate communication (and collaboration) between international fNIRS labs.

To summarise, the HRI research community can greatly benefit from using fNIRS as an innovative tool in embodied interaction studies with humanoid robots. HRI researchers, who use this new methodology, may further profit from improving data sharing infrastructure like the sNIRF file format and data bases such as Openfnirs.

7.5. Future directions

To summarise, lessons learned through the replication crisis in psychology and neuroscience, and the subsequent increased implementation of open science practices, may be of great value to the interdisciplinary HRI research community at large (Belpaeme, 2020). As various researchers have noted, conducting embodied HRI experiments in conjunction with using brain imaging tools is not a trivial challenge – various pieces of hardware and software have to be synchronised to ensure a smooth experimental procedure (Belpaeme, 2020; Perez-Osorio et al., 2018; Strait et al., 2020). Belpaeme (2020) foresees a new future in which experiments will benefit from increased rigor and the field might move towards more transparency in reporting findings – including ‘failed’ replications and null results.

7.5.1. Valuable tools (or: Where to go from here?)

Going forward, researchers working at the intersection of social robotics, experimental psychology and cognitive neuroscience should strive to pre-register their hypotheses, study designs and analysis plans. Several platforms are available for this purpose. Compared to the AsPredicted format, which is designed to cover essential aspects of the experimental design and analysis in nine short questions (e.g.: https://osf.io/ky4b7/), OSF preregistrations allow for more detail and nuance, especially when specifying mixed effects models (e.g: https://osf.io/a5fby). As many researchers have observed, there is an increased uptake of Bayesian analysis methods across all fields (Baxter et al., 2016; Belpaeme, 2020; Shrout & Rodgers, 2018). Although more diverse analytical approaches are commended, researchers also point out that analysts should follow a “principled Bayesian workflow” (Schad et al., 2020, p. 1). For example, Bayesian modelling, which was described in **Chapter 4**, offers the researcher great flexibility regarding the types of models that can be specified and the kind of data that can be modelled. Despite accessible R packages, like ‘brms’, beginners may struggle to navigate across the often-confusing labyrinth of decisions that need to be taken in modelling (Bürkner, 2016; Schad et al., 2020). Here the authors again emphasize the crucial need for specifying models that describe a maximal effects structure already in the pre-registration, to limit later researcher degrees of freedom in the analysis. With greater acceptability of the publishing of null results will come a greater need for reporting evidence for the null. Although the modelling approach used here offers great flexibility, in many cases free programmes such as JASP might be a more beginner-friendly first step in Bayesian analysis (Belpaeme, 2020).

One promising opportunity for researchers at the intersection of these disciplines lies in new article formats (and journals), that place a focus on methodological rigor and meta-scientific perspectives. The vicious cycle of a decreased reward structure to pursue replication projects may be broken by Registered Reports (Chambers et al., 2015) and Registered Replication Reports (RRR) formats (Simons et al., 2014), which encourage large scale collaborative efforts and ensure acceptance at the journal before the results are known (Schimmack, 2020; Shrout & Rodgers, 2018). Adopting these new article formats may contribute to an overall more accurate picture of the scientific evidence base, as it has been recently revealed that the reporting rate of significant results is reduced from about 90 to a mere 50% when the Registered Report format is used (Scheel et al., 2020).

Importantly, these strides to encompass greater transparency in research methods should ideally be accompanied by an acknowledgement and awareness of systematic disadvantages underlying large parts of the existing literature (Dworkin et al., 2020; Pownall et al., 2020; Zurn et al., 2020). In a recent preprint, we argue that just as the open science movement has prompted researchers to adopt more transparent approaches to research, feminist psychology has much to contribute in constructing an equitable movement towards open science (Pownall et al., 2020). One small step could be to include citation diversity statements (Zurn et al., 2020), as implemented in **Chapter 6** of this thesis. Whilst not completely without problems, this new convention may contribute to more transparency about diversity issues evident across scientific disciplines.

7.5.2. Transparent data visualisation

Another important tool for future HRI studies will be to use transparent data visualisation to communicate research findings (Allen et al., 2019). Repeatedly, HRI researchers have lamented issues with the interpretability of psychology and neuroscience findings, noting a missing shared language between disciplines (Baxter et al., 2016; Belpaeme, 2020; Irfan et al., 2018). One crucial factor in efficient communication of experimental findings is the use of clear data visualisations. Currently the most commonly used visualisations across research and the news media remain bar charts, although it has been shown that these visualisations lead to poor decision making when interpreting experimental findings (Newman & Scholl, 2012). For example, Newman and Scholl (2012) found that when they presented bar graphs, participants were more prone to believe that the data were contained within the bars. This visualisation method thus does not offer a good impression of the often-chosen measures of central tendencies. A popular new visualisation method - the raincloud plot – may be a better approach. These graphs depict raw data, distributional information and a boxplot with the median and interquartile range (Allen et al., 2019).

Throughout this thesis, I have used various types of visualisation methods that offer the advantage of “inference at a glance” (Allen et al., 2019, p. 33), including pirate plots (**Chapter 3**), box plots, density graphs of distributions (**Chapter 4**) and raincloud plots (**Chapter 6**). In his chapter on ‘fair statistical communication’, Dragicevic (2016, p. 291) applies “End User Dissatisfaction” (p. 311) as a metaphor. He critically reflects that the field of human-computer interaction (adjacent and interrelated with human-robot interaction, see Figure 1), despite its strong tradition of user experience studies, has adopted visualisations that lead to suboptimal communication of empirical findings. New approaches to graphically representing data may serve as better “user interfaces meant to help researchers in their task of producing and disseminating knowledge, [and] the fields of HCI and infovis can take a head start and show the way to other disciplines." (Dragicevic, 2016, p. 326)

7.5.3. Our future with social robots: beyond 'the social brain'?

Recent years have not only seen a shift in methodological and analytical approaches, but also a move towards incorporating new theoretical perspectives on social cognition (Cross & Ramsey, under review). As these authors argue, the current perspective of using a predominantly anthropocentric approach to investigating interactions with robots may limit the scope of potential questions and might overall stifle progress in this research area. Cross and Ramsey (under review) urge researchers at the intersection of psychology, neuroscience and social robotics to consider a wider, shared feature space between social agents (like humanoid robots) and objects, rather than focusing solely on the commonalities and differences of processing human social interactions compared to interactions with machines. Overall, the authors argue that a more domain-general understanding of human cognition should be adopted, which echoes recent critical reflections by researchers requesting more nuance when parsing 'the social brain' – as it remains to be investigated how the purported specificity for social perception may be represented in the brain (Lockwood, 2020). These recent criticisms can also be seen as a challenge of Chevallier's Social Motivation Theory (2012).

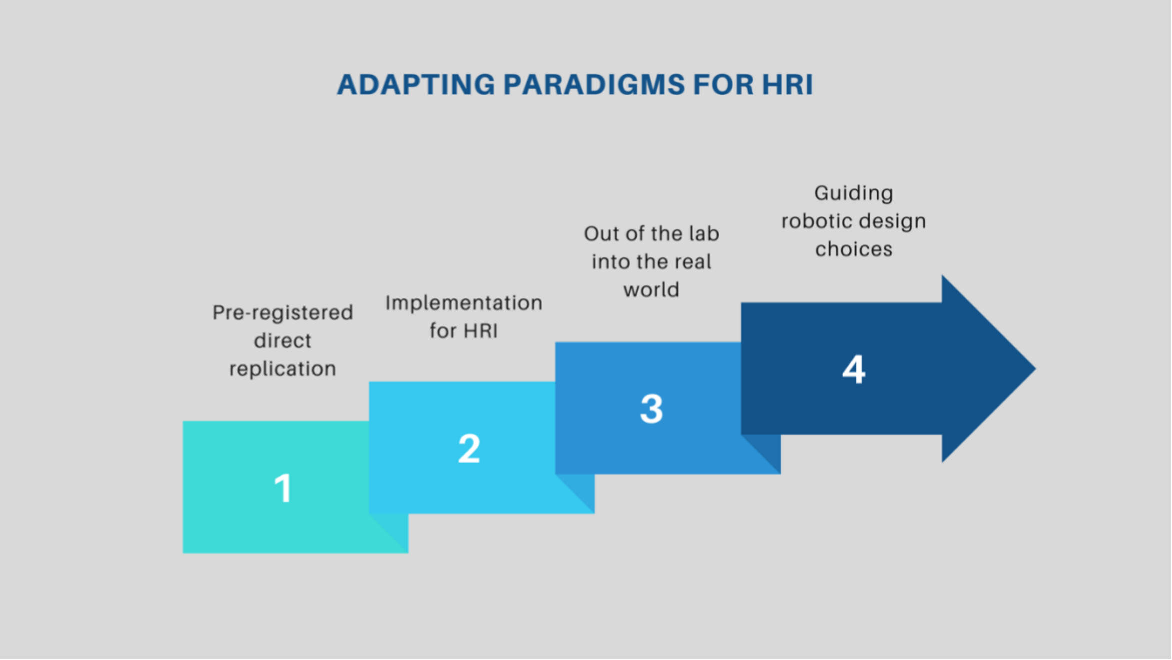

ADAPTING PARADIGMS FOR HRI
Pre-registered direct replication
1
Implementation for HRI
2
Out of the lab into the real world
3
Guiding robotic design choices
4

Figure 27 - Synthesis of the proposed stepwise process.

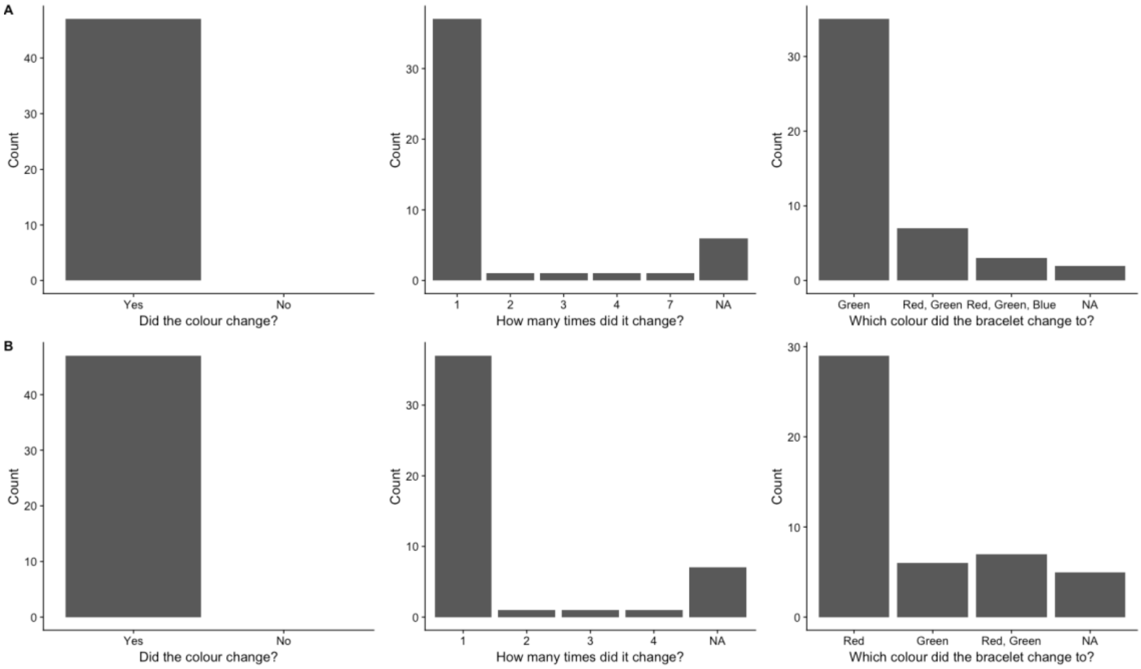
To summarise, the future of adapting paradigms for HRI may include new methodological, analytical and theoretical approaches, that will contribute to our understanding of sharing a social sphere with artificial agents. Through Figure 27, I have attempted to integrate the messages of several researchers working at the intersection of psychology, neuroscience and HRI, as well as one of the main conclusions from this thesis: before conducting studies with humanoid robots, researchers should aim to conduct a (pre-registered) direct replication of the effect first (Irfan et al., 2018), then implement and adapt the paradigm appropriately for (embodied) interactions robots (Perez-Osorio et al., 2018), take these paradigms outside of controlled lab environments into the real world (Henschel, Hortensius, et al., 2020b; Pinti et al., 2018), and finally utilise the knowledge gained to inform the design of social robots (Wiese et al., 2017). These longitudinal and interdisciplinary efforts, as I have argued, are a fundamentally challenging, yet ultimately rewarding undertaking, which may herald a sustainable future for social robots as the helpful companions we envision them to be.

7.6. Conclusions

In this thesis, I sought to adapt methods derived from psychology and neuroscience for the study of embodied interactions with humanoid social robots. Resting on the shoulders of Social Motivation Theory, I have shown that an effect that has already been implemented for some robotic platforms (i.e. the ability to synchronise movements) may have little grounding in empirical truth. Acknowledging the limitations of this null result, I conducted experiments to develop a robust behavioural measure for social motivation, adapting a well-known effect for HRI. Here, the thesis contributes the important insight that this putative behavioural proxy for social motivation may not be easily translatable to HRI, as I failed to replicate the original effect of salient human social cues in this conceptual extension. Taking these mounting findings into account, I proposed that when integrating new methodologies into HRI research, a stepwise process should occur. One emerging methodology, mobile brain imaging, can be used for enhanced ecological validity in the study of social robots outside of the lab. Finally, I introduced a new method to spatially register optode positions in fNIRS studies and compared the fNIRS signal to fMRI via a validated and robust localiser task. Of course, isolated pieces of empirical work cannot be the end of the story on social motivation towards humanoid robots – further replication efforts are needed to challenge or confirm these findings with new robotic platforms and in different experimental contexts. The 2017 workshop ‘The Emerging Social Neuroscience of Human-Robot Interaction’ set the stage for this budding field of research, bringing together researchers from neuroscience, robotics, social cognition and engineering, all with a common interest to leverage advances in social neuroscience to inform and advance HRI. Overall, the findings of this thesis contribute toward this aim, by placing a stronger emphasis on open science methods in the field and contributing to a more realistic picture of the nature of social encounters between humans and robots.

Appendix A Supplementary materials for "No evidence for enhanced likeability and social motivation towards robots after synchrony experience"

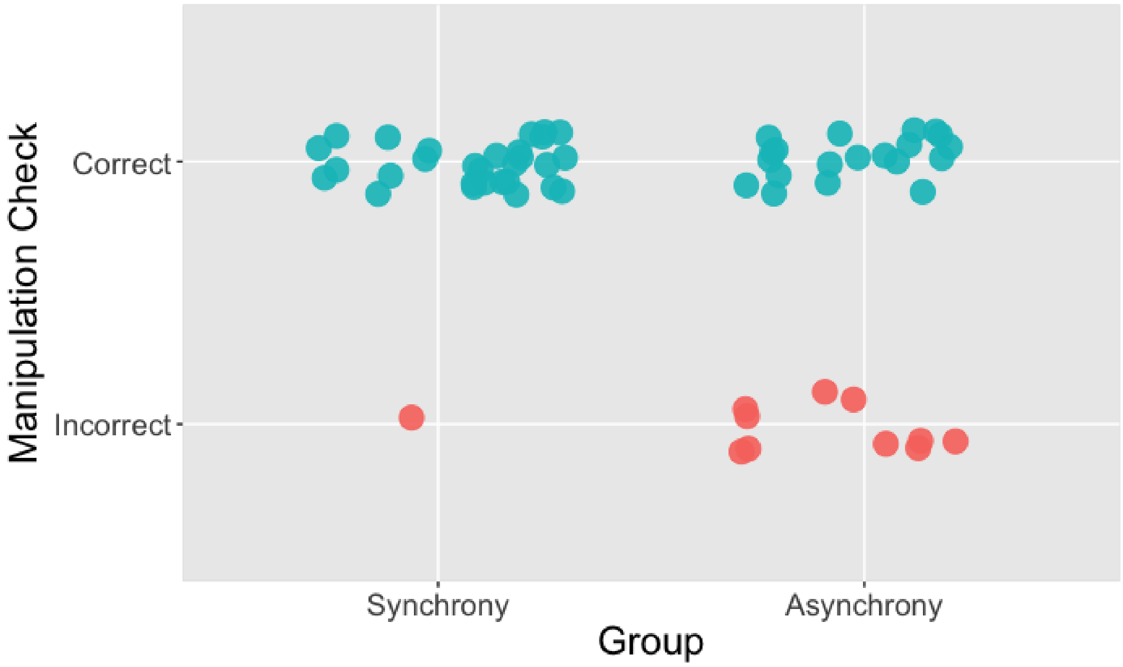
A) Objective manipulation check: LED bracelet colour changes



F1 - Descriptive visualisation of the LED bracelet-based attention check.

Participants were asked to report potential colour changes of the LED bracelet on Pepper's arm. There were two colour checks, one after the first three drawing blocks and one after the final three drawing blocks. Participants first had to report if they noticed any colour change (the correct answer is yes, there was one colour change), then how many changes they observed, and which colour the bracelet changed to. In the first check, the correct colour the bracelet changed to was green, in the second round the bracelet changed to red. Due to technical difficulties with the remote control of the LED lights, it is however not informative to interpret these results beyond the obvious fact that a majority of the participants reported the correct answers on all six checks.

B) Subjective manipulation check



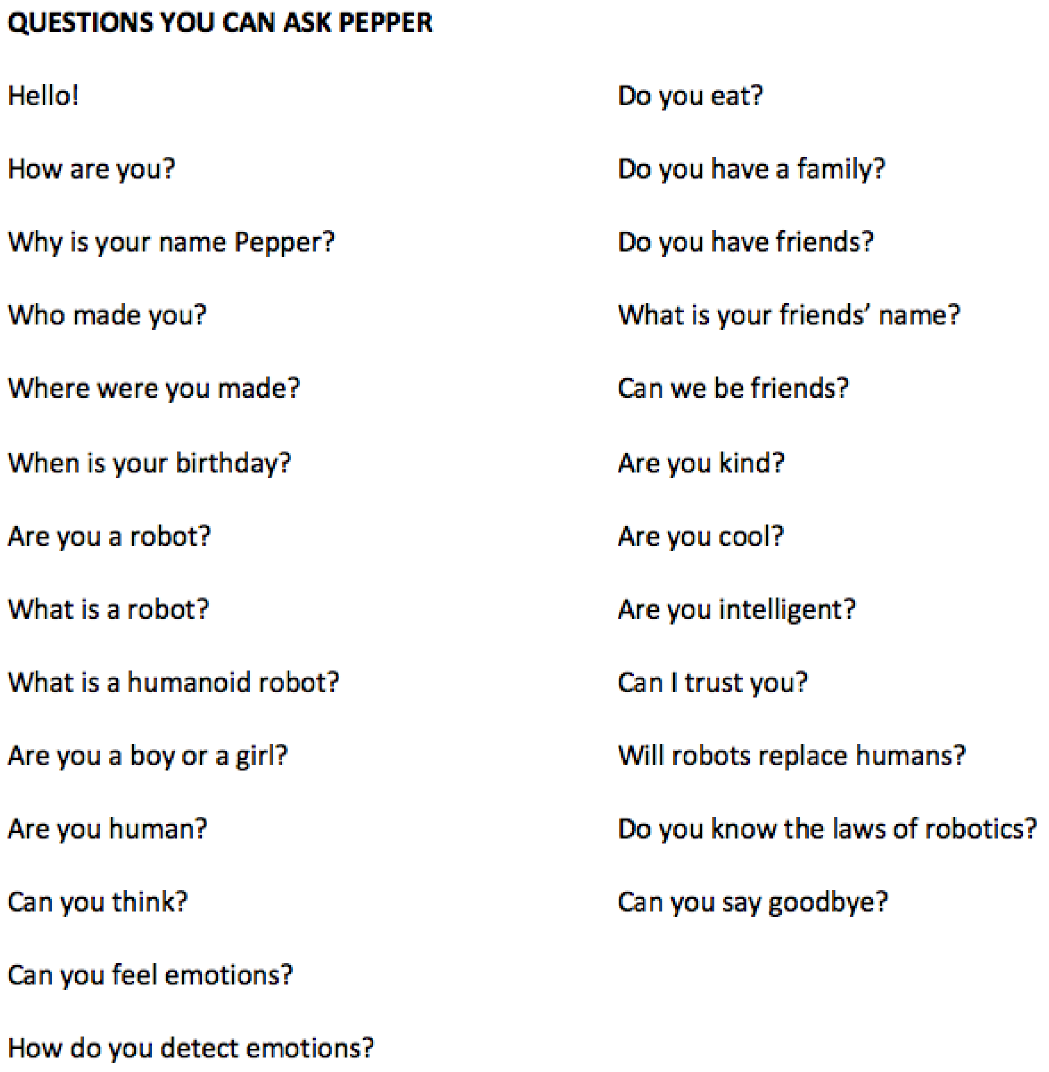
F2 - Descriptive visualisation of the subjective manipulation check.

To probe perceived synchrony, we asked the participants “Did the robot draw … in synchrony with you? …out of synchrony with you?” 10 participants in the asynchrony group reported to have been in sync with Pepper on the drawing task, whereas one participant in the synchrony condition reported to have been out of sync with Pepper.

C) Table specifying the group compositions

T1 - Participant numbers in the planned analysis		
Asynchrony	Synchrony	Total
19	26	45
Participant numbers in the exploratory analysis		
Perceived asynchrony	Perceived synchrony	
20	36	56

D) List of questions participants could choose from



QUESTIONS YOU CAN ASK PEPPER

Hello!

How are you?

Why is your name Pepper?

Who made you?

Where were you made?

When is your birthday?

Are you a robot?

What is a robot?

What is a humanoid robot?

Are you a boy or a girl?

Are you human?

Can you think?

Can you feel emotions?

How do you detect emotions?

Do you eat?

Do you have a family?

Do you have friends?

What is your friends' name?

Can we be friends?

Are you kind?

Are you cool?

Are you intelligent?

Can I trust you?

Will robots replace humans?

Do you know the laws of robotics?

Can you say goodbye?

F3 - List of questions.

The maximum amount of questions participants could ask Pepper was 28 (the two additional questions resulting from participants being able to ask for the second and third law of robotics after Pepper cites the first one. However, since this was a free interaction, some participants chose to either ask zero questions or asked more than 28, in which case we had programmed the robot to be able to answer "I don't know", "Maybe", and "Yes" or "No". Thus, individual participants would end up with a score higher than the number of questions provided by us.

Appendix B Rebuttal for “No evidence for enhanced likeability and social motivation towards robots after synchrony experience”

NB: Reviewers gave consent for their anonymous comments to be shared as part of this thesis. The Media Consent forms remain with the guest editors of the Interaction Studies special issue.

We would first like to thank the reviewers and the editors for their constructive and helpful comments. We very much appreciate that they have taken the time to help us improve this manuscript. We have revised the paper according to their suggestions and have detailed our response in the comments below. Please note that all changes to the main manuscript are denoted in **bold face font.**

The major changes that we have made relate to Reviewer 1’s concerns about the data analysis. We have followed their suggestion to exclude participants from the main analysis, who failed the subjective manipulation check. We have followed Reviewer 1, 2 and 3’s suggestions to provide stronger links to previous literature to justify the task and to illustrate the expected positive effects of interpersonal synchrony on the robots’ perception and behaviour towards it. We have added more points to the critical discussion of the nature of these null results and hope to have addressed all of the editors’ and reviewers’ concerns by doing so. The changes in the manuscript have been marked with track changes. The data and the R analysis script are now openly available via the OSF [link].

Editors’ comments

Comment 1 – We would like to add to the reviewer’s comments that it should be much clearer what the exact relationship is between cognitive neuroscience and psychology is in regard to how these fields have been drawn on for the presented study, it seems that studies on cognition on the level of neurons is today a more independent field of research?

To clarify the link between cognitive (neuro)science and psychology, we cite Wiese and colleagues (2017), who argues that neurocognitive methods can help develop advanced social robots, and Wykowska and colleagues (2016), who elaborate that by using robots in psychological experiments we can learn more about the scope and limitations of human social cognition. These two lines of argumentation are the underlying scaffolding for the present study. Embedding cognitive and experimental psychology, we use the Social Motivation Theory by Chevallier and colleagues (2012) as a theoretical framework, to experimentally test if synchronizing with a robot can improve its likeability and participants' social motivation towards it.

Line 73: "Wiese and colleagues (2017) suggest that the best way to make robots appear more social is to use the toolbox provided by neurocognitive research methods to implement empirically supported behaviors that give "socially awkward" robots better "people skills"."

Comment 2 – Moreover, it would be great to consider having a separate section on lessons learned (instead of including these points in different places in the manuscript) to make it very explicit that there were problems with the design of the experiment.

Thank you for this very helpful suggestion. We agree that these points should be collected in one place. A final section entitled *The Pepper robot as an experimental confederate: lessons learned* (line 352) has been added to the manuscript.

Comment 3 – We also wonder why the problem of "awkward" social robots is paradoxical? It is unclear what this sentence aims to suggest given that it only seems to be a challenge for developers to make robots seem more social.

We agree with the editors that this sentence was unclear. It refers to two papers by Duffy (2004) and by Duffy and Joue (2005): 'The Paradox of Social Robotics: A Discussion'. In encounters with naïve participants, off-the-shelf humanoid robots such as Pepper can still come across as awkward. We believe that making robots appear more social is a team effort to be undertaken by developers and roboticists based on evidence derived from psychological research and user studies. However, to avoid any misinterpretations, the sentence in line 78 has been rephrased to make the message clearer.

Line 77: “But how can we solve the problem of designing truly social robots (Duffy & Joue, 2005)?”

Reviewers’ comments

Reviewer 1

Comment 1 - A methodological drawback, which is also pointed out by the authors, is related to the choice of the synchrony manipulation: Firstly, the authors do not motive their choice for drawing as a suitable task that can induce the experience of synchrony and expected effects of rapport. In fact, a third of the study participants failed to experience the intended feeling of asynchrony during the drawing task with the robot (p 11). Thus, it seems that this manipulation was too subtle to induce the experience of synchrony. Perhaps this could have been avoided by running a pilot study to test the perception of the conditions?

We chose this task based on conceptual and practical grounds that are now described on line 197:

Line 197: “We modelled our task after Hove and Risen (2009). In their study, participants were following a visual metronome (a rising and dropping bar), which resulted in them tapping either in synchrony or out of synchrony with a confederate (Hove & Risen, 2009). Similarly, we used a visual metronome (a small circle moving along a larger circular trajectory) and instructed participants to follow its movement with the pen. The practical reason for choosing this task was that it gave us a high degree of control of the participants’ movement, without explicitly asking them to synchronize with the robot, a potential confound. In the synchrony condition the metronome was linked to the movement of the robot, whereas in the asynchrony condition the robot was moving approximately 2.5 times as fast along the circle shape as the participant. Participants received the instruction from the experimenter that the goal of the task was to follow the moving target as closely as possible and deviate from it as little as possible.”

Comment 2 - Secondly, it seems there was no objective measure of manipulation check. Due to technical difficulties, it was not possible to analyse the answers to the colour change attention test - this would have been a more accurate way to check whether the participants accurately realised whether they were in sync with the robot or not. It would be interesting to report precisely what type of questions were used to check for self-reported perception of synchrony with the robot (i.e. yes/no question, scale).

To probe perceived synchrony, we asked the participants finally “Did the robot draw …

- In synchrony with you

- Out of synchrony with you?”

We agree that the more objective attention check is preferable and have included the colour checks in the supplementary material. Unfortunately, it is impossible for us to trace back in which cases the remote control was not working as expected and in which cases participants simply gave the wrong answers. However, looking at the plots in Appendix A, most of the participants gave the correct answers on all instances of the checks. Thus, we can be confident that they were indeed able to at least see the movement of the robotic arm.

[To avoid duplication, graphs in the supplementary materials are cross-referenced in the rebuttals.]

Figure F1, Appendix A

Comment 3 – Social motivation was assessed using a behavioural measure – the number of questions the participants ask the robot in a free interaction, after completing a drawing task together. As discussed by the authors, this measure was probably “too crude” to reveal the participant’s true motivation to engage with the robot socially. A different type of behavioural or neural measures could have generated more objective findings. Numerous

factors – unrelated to the preceding drawing task might have influenced the number of questions asked by the participants. Some of those factors are discussed in the paper. Prior experience with the robot is another factor that might have influenced the participants' behaviour – someone who is unfamiliar with the robot might be more "curious" to interact socially with it -potentially irrespectively of its behaviour during the drawing task. A free verbal interaction with the robot is a task that involves a different type of robot skills. Therefore, independently of the robot's synchronisation abilities, it is possible that some participants wanted to test the robot's "intelligence" or verbal interaction abilities, and thus engaged in this task.

We concur with Reviewer 1 that the measure we chose to quantify social motivation towards the Pepper robot may have been to crude to pick up on the subtle effects the experience of synchrony might have had on their behaviour towards the robot. However, an advantage of using this measure is that it ensures high ecological validity, as this type of interaction is mainly how users are currently interacting with Pepper. Researchers in human-robot interaction are actively trying to implement reciprocal and synchronous movements into the behavioural toolbox of robots (see: Lorenz, Weiss & Hirche, 2016). However, if for example movement synchrony doesn't affect the quality of the real interaction with a user, this is critical for HRI researchers and roboticists to know. Indeed, a more objective measure, such as neural activity would have been desirable. We are currently working on developing more objective measures of social motivation towards robots, as there appears to be a scarcity of them available to HRI researchers. Following the suggestions of Reviewer 1, we have included 'prior experiences with the robot' as one of the factors that could have played a stronger motivational role than the preceding experience of synchrony/ asynchrony in the manuscript.

Line 327: "In addition, previous experiences with the robot might have influenced their behavior, with participants lacking any experience perhaps showing stronger curiosity to interact with Pepper or a lack of familiarity affecting the mind perception of the robot (Müller et al., 2011)."

Comment 4 – Finally, my biggest concern is related with the data analysis. Since the goal of the study was to investigate whether synchrony has an

effect on social motivation towards the robot, I would expect that participants who did not perceive the synchrony manipulation as intended, would be excluded from the analysis (manipulation check). Although the authors claim to take this into consideration (p 9, line 198), the data analysis reporting is difficult to follow. The "Original group split" results, as well as Figure 2, do not take into account the manipulation check - and are therefore not conclusive (by mixing the ratings of participants who failed to perceive the synchrony effects and participants who accurately perceived them it is not possible to draw any conclusions about the effects of synchrony on the dependent variables).

We thank Reviewer 1 for bringing to our attention that aspects of our data analyses were difficult to follow. We can see how some confusion could arise following our line of argumentation in the paper. Based on prior work by our group, as well as other research teams, on the role of participants' beliefs about artificial agents' behaviour (c.f., Klapper et al., 2014; Cross et al., 2016; Liepelt & Brass, 2010, Wiese et al., 2012), we were interested in exploring participants' top-down perception of synchrony in addition to the actual synchrony manipulation. This is why we also split the data based on participants' synchrony beliefs. Since we can be fairly confident that a majority of participants' attention was on Pepper's arm (due to the additional colour change detection task, added in the supplementary material of the manuscript, **Appendix A)**, we explored whether the subjective experience of participants would play a more important role than objectively manipulated synchrony. To address the concerns of Reviewer 1, we have replaced the first analysis with a new analysis following Reviewer 1's suggestions: we have excluded all participants who failed the subjective manipulation check. However, we have kept our second exploratory analysis in the manuscript, as we consider the resulting 'perceived group' split still very interesting, and potentially valuable for future studies to pursue. In addition to the corrected results section and figures in the manuscript, our analysis script and data are now available via the OSF for any interested researchers to explore further.

Line 159: "11 participants were excluded, as they failed the manipulation check of correctly perceiving synchrony or asynchrony."

Comment 5 – In Figure 3, "Individuals, who were in the asynchrony condition, but reported to have been in sync with Pepper were combined with those, who were objectively in sync with the robot.". It is unclear why the authors follow this approach. Ideally one would expect to see the graphs showing only the ratings of participants who passed the manipulation check and were thus objectively in/out-of-sync with the robot.

We have amended the data analysis following the suggestions in the previous comment, so that Figure 2 now reflects this case. However, as illustrated above, we have kept the exploratory analysis for the perceived groups and thus Figure 3 remains in the manuscript. To avoid confusion of how the final groups for the analysis are composed, we have added a table in the supplementary materials (**Appendix A**) and added this information in the caption of Figure 5 and Figure 6.

Table T1, Appendix A

Comment 6 - Taken together, the points mentioned above render the experimental results of this study rather weak and inconclusive. Even though the lessons learned from this study are interesting and relevant in terms of "What Could Go Wrong during HRI studies", the overall impact is limited. In revising this paper, the authors should justify methodological choices and focus on improving the data analysis, as well as the clarity of the presentation of the results.

Following the suggestions of Reviewer 1, we have clarified our methodological choices, data analyses, and figures.

Reviewer 2

Strengths

a) **Well-organized paper. Very straightforward use of a couple key measures (though they could be explained better and justified better).**

b) **Great job of talking about the results section, the analyses used, and displaying your data!**

c) **The authors seemed quite intentional about making sure they included only participants for whom the study worked well. Keep it up :-)**

We thank Reviewer 2 for their positive feedback. There might have been a slight misunderstanding regarding how we excluded participants. Though this study was unfortunately not pre-registered, we excluded participants based on a pre-defined set of criteria. The rationale behind the exclusion criteria is as follows:

- In the case of participants who deviated significantly from the metronome, the manipulation would not work as desired, thus those with a large error rate were excluded.

- Those participants with missing responses on the crucial Godspeed questionnaire were excluded as well.

- Two more participants were excluded because despite our recruitment criteria, they reported studying computer science. Our rationale behind this was that computer scientists might be more sceptical towards the robot, and in the free interaction period might want to test the robot's inbuilt AI, instead of focusing on the social aspect of the interaction.

- One participant reported a diagnosis of ASD, which, based on previous literature on altered social motivation in individuals with ASD (Chevallier et al., 2012; Chevallier et al., 2013), we had also defined as an exclusion criterion.

- Following the suggestions of Reviewer 1, for our main analyses we also exclude all participants who failed the subjective manipulation check as well.

Weaknesses

Comment 1 - Missing some pertinent information in the literature review regarding studies on synchrony with robots. I include recommended studies to look at the detailed comments.

In agreement with the comment made by Reviewer 2, we now include additional citations based on the literature recommendations given to us. We believe the revised manuscript now covers the key aspects of the relevant literature.

Comment 2 - Further, the study does not take into account other factors of importance without which synchrony with the robot will have no effect. It is unclear if participants thought the robot had any intentionality or that the robots motion had any meaning for them, without which, participants have already been shown to have no effects of synchrony (Oberman, McCleery et al. 2007, Press, Gillmeister et al. 2007, Wiese, Wykowska et al. 2012).

We thank Reviewer 2 for raising this issue and for drawing our attention to these studies. We address this problem starting from line 303 in the discussion section. We have now included the citation of the study by Wiese and colleagues (2012), which is in line with the point we are making here:

Line 303: "In addition to the potential necessity of adaptivity in synchronous interpersonal movement, Lorenz, Weiss and Hirche (2016) argue that in order to reap the benefits of synchrony in social interactions with robots, the human interaction partner needs to attribute a mind to the robot. This idea is consistent with research by Wiese and colleagues (2012), which shows that top-down beliefs about an agent's intentional stance can influence basic attentional mechanisms. Even though we assessed trait negative attitudes towards robots, we did not include a self-report or behavioral measure of mind attribution. While Pepper introduced itself before starting the drawing task, it remains unclear how much independence and intention the participants attributed to the robot."

Comment 3 - How did the authors choose their synchrony task? The finding of no effect of synchrony on liking of/talking to the robot would be more interesting if there were a condition in which synchrony with the human under the same circumstances did increase liking/talking to the human. This is something that the authors can do, and I recommend running parallel human conditions for future studies if they follow up on this paradigm.

Regarding the nature of the task, please see how we addressed comment 1 of Reviewer 1. We agree with Reviewer 2 that the study would be even more convincing if it included a positive control, i.e. a human-human condition that provides evidence for the success of the manipulation. We aim to include this in future studies following up on this one and mention this limitation/future direction on line 338 of the revised manuscript.

Line 338: "Future experiments should further include a positive control to ensure the synchrony manipulation works as expected in human-human interaction and additional loops of control to ensure that the synchrony manipulation is sufficiently immersive and salient."

Comment 4 - Incomplete method section makes it unclear if the lack of effect of synchrony was because there is no effect or because the experimenter treated the robot like a thing (in addition to the concerns in the above bullet point). It will also help to justify the question asking measure you chose using previous literature.

In the methods section, we describe how the robot introduces itself (as a member of the University research department) in the experimental procedure. Experimenters were encouraged to avoid the use of gendered pronouns and instead referred to the robot as 'Pepper' or 'the robot'. Our measure of social motivation, which relied on how many questions participants chose to ask Pepper from a list, was custom made for this experiment, and does not directly relate to a similar measure that has been previously used in human-robot interaction studies. Our rationale for choosing this measure was that human-robot interactions with Pepper in real life are usually characterized by these question-answer dynamics, so we chose this measure to gauge whether our manipulation would have an effect in a relatively natural scenario (see response to comment 3 of Reviewer 1).

Comment 5 - Line 23 "Positively influences likability and prosocial behavior towards that individual" sounds strange

This sentence has been rephrased (line 27):

“A wealth of social psychology studies suggests that moving in synchrony with another person can positively influence their likeability and prosocial behavior towards them.”

Comment 6 - Around line 71, you assume that adaptive behavior is similar to synchrony, but you don't explain why.

Conceptually, there are two forms of synchrony that are discussed in the literature. One refers to orchestrated synchrony, i.e. synchrony that is induced by following a shared metronome (the framework chosen for this study), while the other form of synchrony is naturally emerging and requires adaptive movements from each agent. Shen and colleagues (2015) wanted to emulate naturally emerging synchrony and equipped their robot with an information distance algorithm, designed to promote emerging synchrony between human and robot. Hence, we do not assume that adaptive movements are equivalent with synchrony but are important components of naturally emerging synchrony.

Comment 7 - You are missing a section talking about studies that have already examined synchrony (not just adaptive behavior) with robots. I recommend the following articles:

As we explained in our answer to the above comment, the two studies by Mörtl and colleagues (2014) and Shen and colleagues (2015) are indeed focused on synchrony with robots and even go one step further by trying to emulate the natural occurring synchrony we can observe between humans in everyday life (for example, synchronized clapping at concerts). However, we appreciate that many other articles have discussed human-robot synchrony as well and have studied the articles suggested below. The papers by Oztop, Franklin and colleagues (2005), Kilner, Paulignan and team (2003) and Sartori et al (2011) measure automatic imitation and not interpersonal synchrony. We thank Reviewer 2 for recommending additional references – where relevant we have included them in the paper (see our answer to comment 2 and line 318 in the manuscript).

Comment 8 - It is not clear why participants were excluded for studying computer science. If you exclude the students, should you also exclude psychology students who might be able to guess your purpose?

This decision mainly relates to their more comprehensive knowledge of robotics and artificial intelligence. We expected that they would approach the interaction with the robot very differently to a naïve participant. Furthermore, all of the included Psychology participants were naïve to the purpose of the experiment. Indeed, they repeated the cover story they were told. We explained to them in the beginning of the experiment, that we were interested in investigating how the presence of a robot might affect their performance on a task. We would have removed participants that were able to guess the purpose of the study (but this was not necessary). Please see response to Comment 13 of Reviewer 2.

<u>Line 166: "Participants were naïve to the goal of the experiment."</u>

Comment 9 - Studies indicate that answering demographic information first changes the way the participants respond to experimental protocol. I recommended future studies, you ask them demographic information last.

We were not previously aware of this and thank Reviewer 2 for this recommendation, which we will happily follow in the future.

Comment 10 - When you asked them about "trait attitudes toward robots," - a couple of questions arise. What do you mean by trait attitudes? Has someone validated a scale for trait attitudes? What questions were asked?

When talking about trait attitudes towards robots, we use this as a qualifier to distinguish it from attitudes towards robots that arise due to the state of a situation. We used the NARS (Negative Attitudes towards Robots Scale) by Nomura and colleagues (2006) (in the English version by Syrdal, Dautenhahn et al., 2009). These references have been added to the paper (line 190/191). The scale has been widely used in the field and has been validated in different languages (for example, Picarra et al., 2015).

Comment 11 - Second, the measure that you said you use is that they were responding regarding attitudes about "they" robot (line 141). If they hadn't met the robot yet, how did they do this? What did they know about the robot when they were answering this question?

The ‘they’ here refers to the participants, who were asked to fill out the questionnaires before the start of the task. The Negative Attitudes Towards Robots Scale asks about negative feelings about situations of interactions with robots, the social influence of robots and the negative attitude towards emotions in interactions with robots (Nomura, 2006). Thus, it asks about robots in general, and not the Pepper robot specifically. This is why we refer to it as ‘trait’ negative attitudes of the naïve participants, before they met the robot. We did not enquire what they knew specifically about the Pepper robot before introducing it. However, it is unlikely that our participants had met a Pepper robot previously, as we were the first lab at the University to conduct an experiment with it.

Comment 12 - How did you treat the robot? Did you call it by name, treated like a human, or did you treated like a thing? Some of the above studies cited indicate that if people don't think the robot is intentional, synchrony won't matter. If the experimenter doesn't treat the robot as a human, people likely will not perceive it as intentional.

As we have addressed this point already above (Reviewer 2, comment 4), we only want to briefly explain that the robot was treated as a supposed ‘member of the research department’ and was referred to by the experimenters as ‘Pepper’ or ‘the robot’.

Comment 13 - What was your cover story? You emphasize the importance of experimenter bias, which is great! - but participants might also start guessing your purpose if you don't have a solid cover story and they were made to go in sync with the robot.

We agree with Reviewer 2 that a plausible cover story is very important. We informed participants that we were interested in investigating the effect of the presence of a robot on the performance of a (drawing) task. No participant raised suspicion and all of them remained naïve to the true purpose of the experiment. See response to comment 8 of Reviewer 2.

Comment 14 - What questions were on the paper that they picked up (line 181)?

We have added the list of questions to the supplementary material (**Appendix A**).

Comment 15 - In line 208, P only goes out to one decimal place. Keep it consistent please.

The p-value we found was *.6037*, so we had rounded to *.6*. This has now been amended to *.60*.

Comment 16 - Figures 2 and 3 look great! Very informative.

We agree with Reviewer 2 that the pirate plots generated with the R yarrr package give a great overview of the data, since they show raw data, central tendencies and densities, and the 95% highest density intervals, thus combining raw, descriptive and inferential visualisation. This information has been added to the figure captions on page 11.

Comment 17 - When participants were split by perceived synchrony, how many people were in each condition?

We have addressed this question by adding a table in the supplementary material (**Appendix A**). The information on how many subjects were in each group has also been added to the figure captions on page 11.

Comment 18 - The paragraph starting in line 224 is great, love it! The paragraph after it is also very important. I wish this were earlier and that the study could have included this information.

We feel like the position of this point in the discussion of the findings is appropriate, given that we are considering here explanations for why we are observing null results using this particular experimental manipulation. In our next studies, we will keep in mind to consider perceived intentionality and add a measure to capture it.

Comment 19 - The information in paragraph starting on line 259 is great! I think it should be included in the method section because it relates to how you actually ran the study.

We respectfully disagree with this suggestion, especially as it was suggested by the editors to collate the lessons learned into one section, which seems appropriate at the end of the paper, following the logical flow of ‘what have we learned from the human-robot interaction study as experimental psychologists’.

Reviewer 3

Comment 1 - Firstly, at times I missed a clear link between the design and the literature, of which I gave the most prominent examples under the minor comments below.

We thank Reviewer 3 for drawing our attention to this problem. We hope that Reviewer 3 finds that the revised manuscript makes the link between the design and literature far clearer, as the other 2 reviewers raised related points which we have addressed in this revision (Reviewer 1: comment 1,3; Reviewer 2: comment 2,3,7,10). In our specific answers in response to Reviewer 3’s points below, we provide more details as to how we have changed the manuscript to reflect this.

Comment 2 - My other main concern is as follows. I believe strongly in the publication of null-results, as long as the study has a scientific contribution, which this paper clearly offers. However, I worry that it might also be driven by the experimental design or by a lack of statistical power. In terms of experimental design, I specifically wonder if the experience of synchrony was too subtle (l.222) because the immersion was not deep enough due to mostly technical constraints (such as robot movements, its screen turned off, and the physical distance between the robot's pen and its screen). Perhaps a brief summary of the manipulation check that was carried out would be informative.

We thank Reviewer 3 for raising their concern regarding the depth of the synchrony immersion. We have addressed Reviewer 1 and 2's concerns regarding the two manipulation checks we carried out and have subsequently added two figures in the supplementary materials, visualising the objective attention check and the check for the participants' subjective impressions of synchrony (Appendix A). While analysing the objective attention check was not informative due to technical difficulties with the remote control that changed the colours of the LED bracelet, we can see that most participants were able to report the colour changes correctly, so we can be fairly confident that they were attending to Pepper's arm movements (also see: comment 2 of Reviewer 2). Future studies might include either more loops of control to ensure that the depth and saliency of the synchrony experience can be quantified or could include a more natural manipulation, based on emerging synchrony between the interacting agents.

Comment 3 - Additionally, was the task analogous to one used in the reported human-human studies? Since the metronome is quite an exogenous cue, perhaps social feedback/interaction is not relevant to the participant's mindset, as they are simply carrying out the task. If other studies used a similar design, please report it. If not, discuss it in more detail; as it would enhance the issue touched upon in the paragraph starting at l.224.

We have addressed the underlying motivation for choosing a paradigm that orchestrates synchrony via a shared metronome in response to Reviewer 1 and 2's concerns (Reviewer 1, comment 1). We have added the information in the manuscript that the task design was modelled on the seminal study by Hove and Risen (2009), who used a similar visual metronome to synchronize finger tapping between their participants and the confederate.

Comment 4 - About statistical power: as far as I understood, there were 6 blocks of 4 trials (=24 trials) in total; or are the 4 repetitions part of the same trial? I realise that no variables were measured from this interaction itself, so the issue I raise here is not one of measure repetition; but I do wonder if it ties back to the immersion aspect. Namely, relative to the duration of the entire experimental session, the interaction was quite short. I feel like a short sentence or two addressing this in the discussion would make this information more transparent.

There seems to have been a slight misunderstanding concerning how the blocks are composed: Each drawing block consists of the participants following the moving dot 4 times around the circle. In total we have 6 drawing blocks, each followed by a break, which amounts to 24 trials (=24 drawn circles). We concur with Reviewer 3 that this indeed led to a rather short immersion into synchrony with the robot, however, this is directly modelled on the short synchrony interventions reported in the human-human interaction literature (Hove & Risen, 2009; Cross, Wilson, & Golonka, 2016). We have added a sentence in the discussion addressing this issue.

Line 292: “Given this finding, it may be that the experimental manipulation of synchrony was either too subtle or too short to fully immerse participants in the experience and to produce the hypothesized beneficial effect on rapport between synchronizing agents.”

Line 338: “Future experiments should further include a positive control to ensure the synchrony manipulation works as expected in human-human interaction and additional loops of control to ensure that the synchrony manipulation is sufficiently immersive and salient.”

Comment 5 - In summary, these concerns can mostly be simply addressed in the discussion section, to provide a more critical evaluation. Other than that, I read this manuscript with great interest and I really recognised a number of issues myself (for example, we recently had a near-ceiling effect of likeability on the Godspeed questionnaire as well: people just like humanoid robots!). I particularly enjoyed the recommendations toward the end. Great work in general!

We thank Reviewer 3 kindly for their feedback.

Comment 6 - Please describe the task in more detail. Were participants meant to trace the visual metronome? I found this unclear.

We thank Reviewer 3 for drawing our attention to the fact that the task description was unclear in the manuscript. We have added an additional sentence of explanation in the Experimental Procedure to make clear that the participants were informed that it was the goal of the task to follow the moving dot (=visual metronome) as closely as possible and too make as little mistakes as possible:

Line 207: “Participants received the instruction from the experimenter that the goal of the task was to follow the moving target as closely as possible and deviate from it as little as possible.”

Comment 7 - Compliments on being so straight-forward about selecting the final pool of participants, which appears analogous to the 21-word statement by Simmons, Nelson and Simonsohn, 2012. However, to complete this, could you provide a sample-size justification?

This experiment was set up as an initial proof of concept study. Our sample size results from our motivation to recruit the highest number of participants in a limited amount of time (i.e., before our lab moved in the spring of last year).

Line 153: “. We aimed to recruit the highest number of participants within the testing period (February to April 2018).”

We now also include a data statement.

Line 150: “*Data statement.* We report all measures in the study, all manipulations, any data exclusions, and the sample size determination rule. The data and the R analysis script are publicly available via the OSF [link].”

And we critically discuss the fact that the sample size might have been too small to detect our effects of interest:

Line 341: A final limitation we would like to highlight is the fact that given the rather high number of participants we had to exclude, the sample size may have been too small to show the expected small to medium effect size of a synchrony manipulation on perception of and behavior towards the robot.

Comment 8 - The goal of the study (paragraph starting at l.115) could do with more explicit linking to the previous paragraph(s) to clarify these links to the

reader. I suggest something along the lines of: "likeability, analogous to the findings by Chevallier et al. (2012)" and "increased number of questions during a subsequent free interaction as a measure of prosocial behaviour".

We agree with Reviewer 3 that this would clarify the links to the previous paragraphs and have added these references at the end of the introduction section.

Line 140: “We hypothesized that moving in sync with the robot would improve its likeability, analogous to the findings of Lehmann and colleagues (2015), and, based on Chevallier’s social motivation theory, would also increase the motivation to interact with the robot, as measured by the number of questions participants chose to ask the robot during a free interaction.”

Comment 9 - Similarly, I would add a "see experimental procedure" after the "losing the metronome" statement in l.123; as the metronome has not been introduced yet at this point.

We thank Reviewer 3 for drawing our attention to this, we have added a reference to the experimental procedure.

Line 154: “Four participants were excluded from further analysis due to large error rates (losing the metronome more than 30 times, see experimental procedure below) on the task, and four more had to be excluded due to missing data on the questionnaires.”

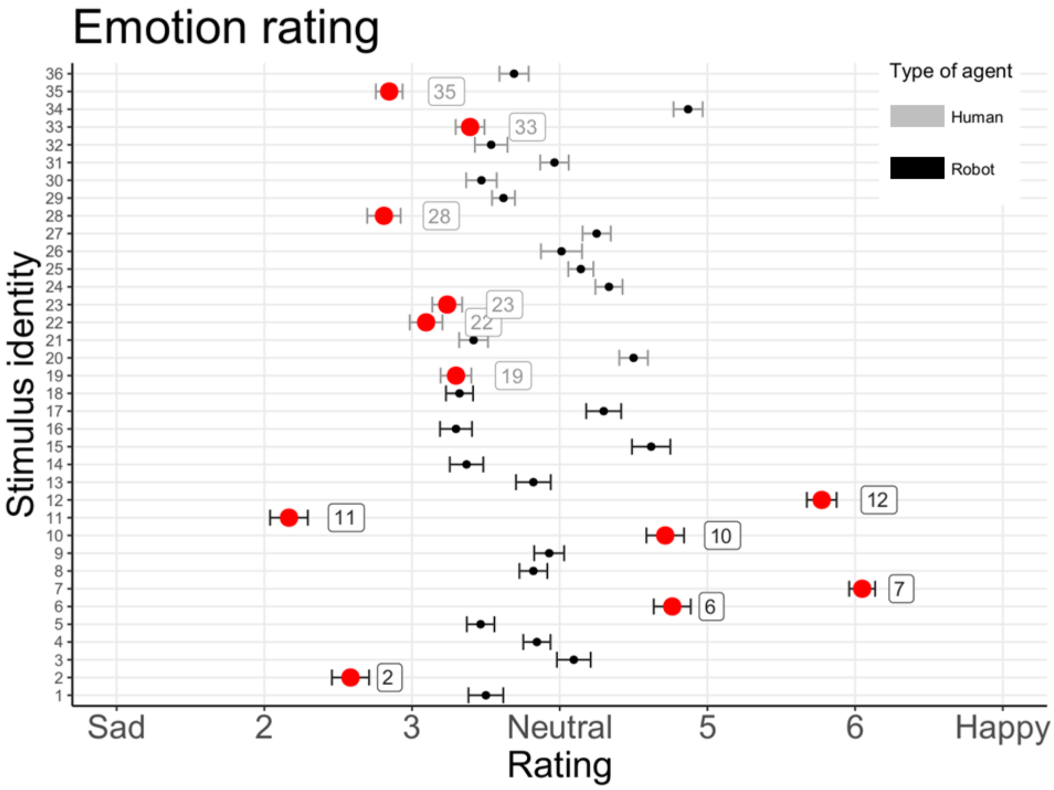
Comment 10 - Would you be able to provide open data plus a link to it in the manuscript?

We have uploaded the data and the R analysis script, as well as the html output file to the OSF and have made the project available to the public. Thank you for your helpful comments.

Appendix C Supplementary materials for “Faces do not attract more attention than non-social distractors in the Stroop task”

A) Emotion rating (online validation study)

Prior to study 1, we ran an online stimulus validation study to ensure that the faces would receive comparable ratings in perceived emotionality. At the time of this online validation study (June 2018), we had not yet added the third control condition (pareidolic faces), so only the emotional content of unique human and robot faces were rated. Furthermore, as this was the first set of stimuli, there were less unique images (12 images per condition) compared to study 2 (24 images per condition). The validation experiment was presented in Jisc Online Surveys (formerly Bristol Online Surveys). Participants rated 18 unique robot and 18 unique human faces (male and female) on a bespoke semantic differential scale between ‘1 – sad’ to ‘7 – happy’. ‘4’ was considered ‘neutral’, for the purpose of the analysis. The scale was made for this study. 84 participants (age: M=34.67, SD=11.77) completed the rating study.



F4 - Emotion ratings.

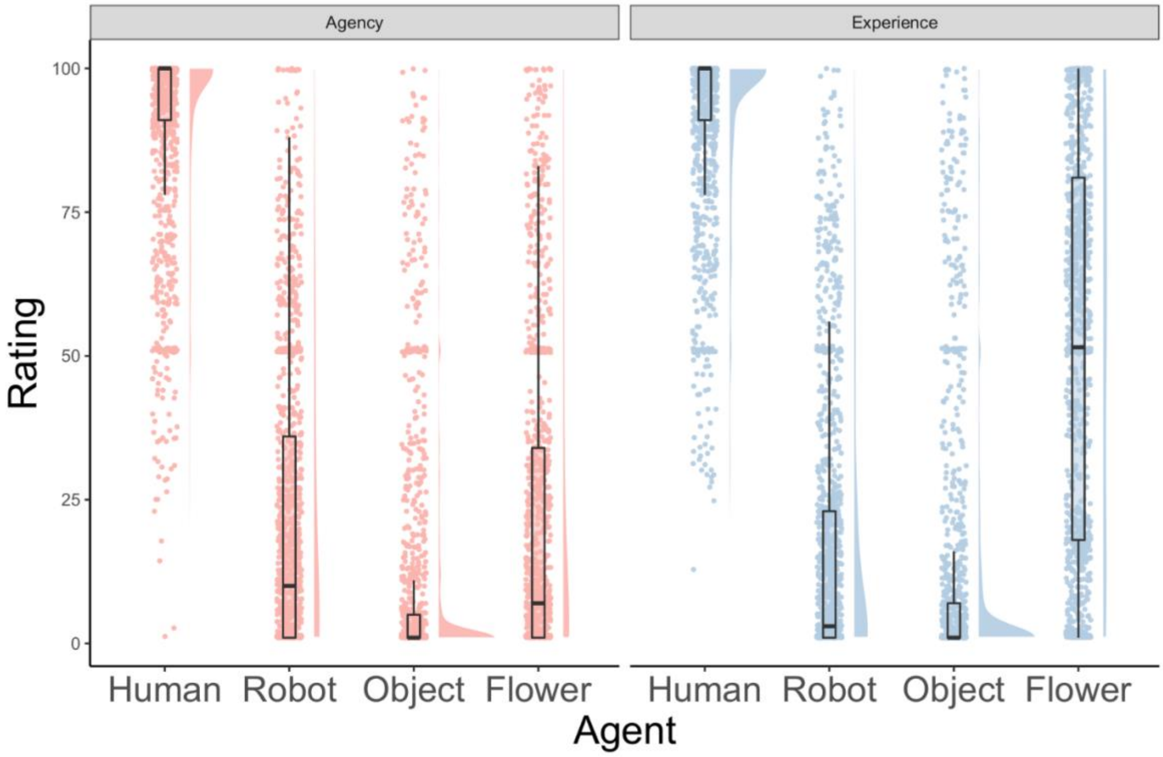
The bold dots represent the mean rating scores for each image, and the bars represent the standard error. Red & labelled points indicate that those images were excluded.

Most participants were female (*n*=64) and most reported never having interacted with a robot before (*n*=61). The participants were recruited via advertisements on social media.

While the two groups didn't differ in mean ratings at first glance (human faces: *M*= 3.69, *SD*=1.08, robot faces: *M*=3.97, *SD*=1.39), ordinal logistic regression with the 'ordinal' package (Christensen, 2019) suggests that human faces were rated more negatively (estimate = -.37, SE= .07, p<.001). Following this result, we inspected the mean ratings of the individual stimuli visually and discovered that the robotic faces were rated much more variably than the human faces. As the stimulus exclusions were costly (i.e. the time and effort to replace and re-process all images), and we had to work towards keeping at least 12 unique images within the pool of 18 images in each condition, we removed the strongest outliers in the robot condition (robots 2, 6, 7, 10, 11, & 12) and removed those human faces that were rated more negatively than the average, with 3 male and 3 female faces each (19, 22, 23, 28, 33 & 35). While the procedure we followed is not optimal (limited by the time and stimulus availability constraints), we gained valuable insights. The robot faces we considered “neutral” upon selection were, in fact, perceived not as unambiguously neutral by the raters, and despite selecting human faces from the neutral condition of the Radboud Faces Database (Langner et al., 2010), they were perceived slightly more negatively than the midpoint of our scale.

B) Agency and experience ratings (Experiment 2)

In conjunction with study 2, we included a second survey component, which required participants to rate all the unique (un-mirrored) images on *agency* and *experience*. Again, we used a bespoke rating scale modelled on the conceptual ideas of Gray, Gray and Wegner (2007), who define agency as the ability to plan and act and experience as the ability to sense and feel. The images and rating scales were presented via FormR (Arslan, Walther & Tata, 2019) and the one-sentence definitions of agency and experience were presented below each image. Fifty-one participants (the same sample as in Study 2) rated 96 unique images on agency and experience (24 per category, 4 categories). The three inbuilt attention checks (for example: "Did the last image show a) an objects or b) a human?" were all answered correctly by all subjects.



F5 - Agency and experience ratings (Experiment 2).

The agency and experience ratings of the 4 stimulus categories: human faces, robot faces, pareidolic faces (objects) and flowers. There is a clustering at the midpoint of the scale, which can be explained by the fact that the starting point of the rating scale was always at 50.

A within-subjects ANOVA conducted with the R package {ezAnova} suggests that there is a main effect of agent: *F*(3, 150) = 189.71, *p*<.001. Mauchly's test for sphericity was significant, thus the assumption was violated (W) = 0.37, p <

.001). The Greenhouse-Geisser estimate of sphericity ($\varepsilon = 0.74$) was used and the corrected p-value remains significant ($p<.01$). Upon inspecting figure 5, it appears that humans were rated highest on agency and experience, robots were attributed some agency and little experience, pareidolic faces rated lowest on both dimensions of mind and surprisingly there was a large spread of ratings for the ability of flowers to sense and feel. This satisfies our internal criterion for ‘category difference’, to ensure that each of the faces were sufficiently distinct.

C) Pre-processing of response times

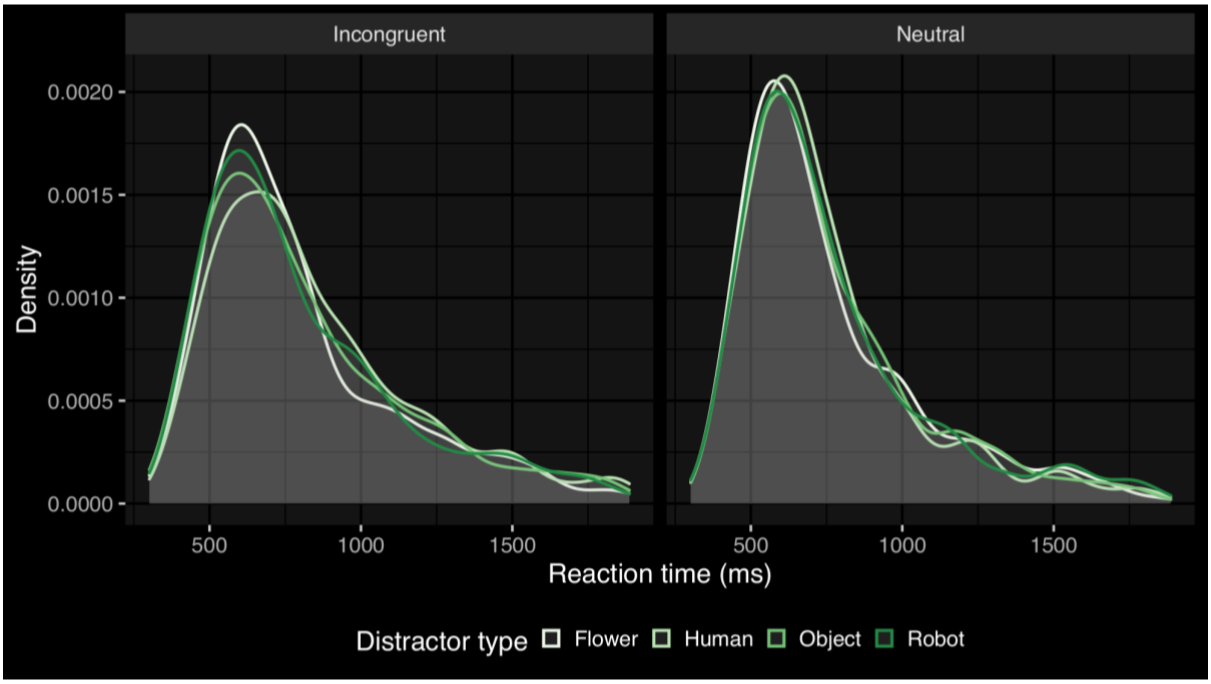
Table T2 lists the number of remaining trials per experimental condition after applying our pre-registered trimming criterion in Experiment 1. Table T3 depicts the mean reaction times using our pre-registered standard deviation criterion and Figure F6 and F7 depict density plots of the response times for each of the pre-processing methods. Applying the suggested standard deviation per participant criterion resulted in overall slower reaction times (Table T3), compared to the original analysis (Table T2). As the pattern of results nonetheless looks similar to what we originally reported, and in the interest of adhering to our pre-registration, we decided to keep the current results section of Experiment 1 as it is.

T2 - Experiment 1: Number of remaining trials per experimental condition after reaction time trimming using the pre-registered standard deviation criterion. 192 trials, 39 participants.

Condition	**Total number of trials**	**Trials remaining**	**% of trials remaining**
incongruent_human	936	847	90.5
incongruent_robot	936	835	89.2
incongruent_object	936	857	91.6
incongruent_flower	936	829	88.6
neutral_human	936	883	94.3
neutral_robot	936	870	92.9
neutral_object	936	883	94.3
neutral_flower	936	878	93.8

T3 - Experiment 1: Mean reaction times (in ms) per condition using the pre-registered standard deviation criterion.

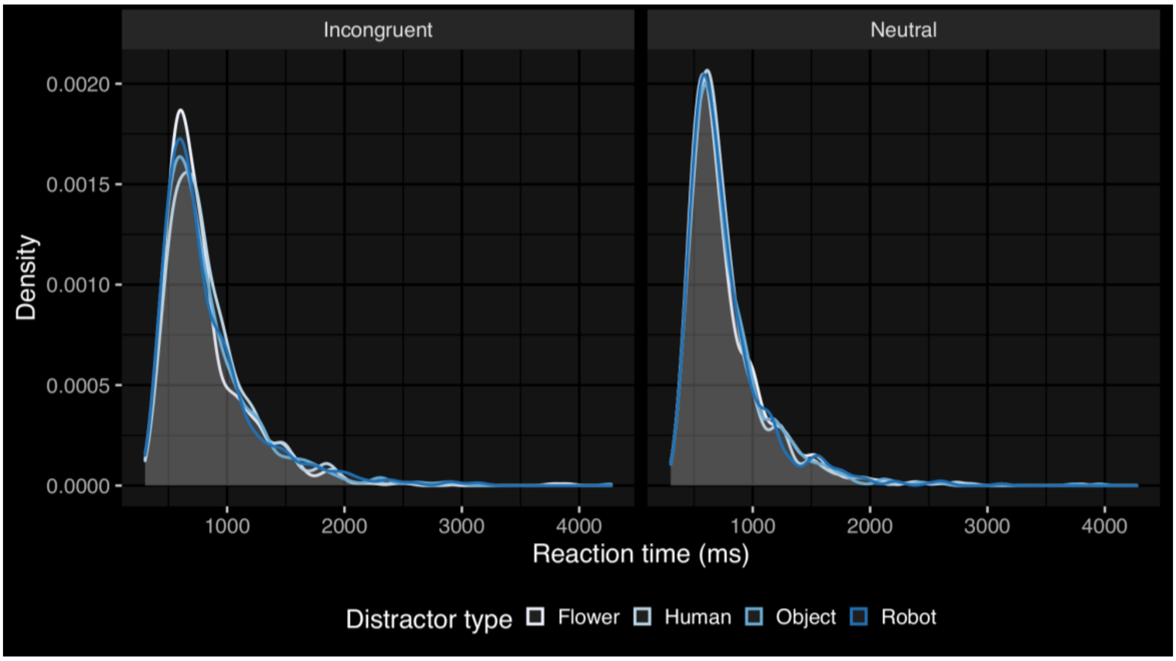
	Distractor			
	Human	Robot	Object	Flower
Incongruent target	843	807	815	796
Neutral target	753	768	763	760



F6 - Density plots for reaction times (ms) in Experiment 1 with the pre-registered standard deviation trimming criterion.

T3 - Experiment 1: Mean reaction times (in ms) per condition using the standard deviation per participant criterion.

	Distractor			
	Human	Robot	Object	Flower
Incongruent target	840	833	833	809
Neutral target	769	770	780	776



F7 - Density plots for reaction times (ms) in Experiment 1 with the participant-sensitive standard deviation trimming criterion.

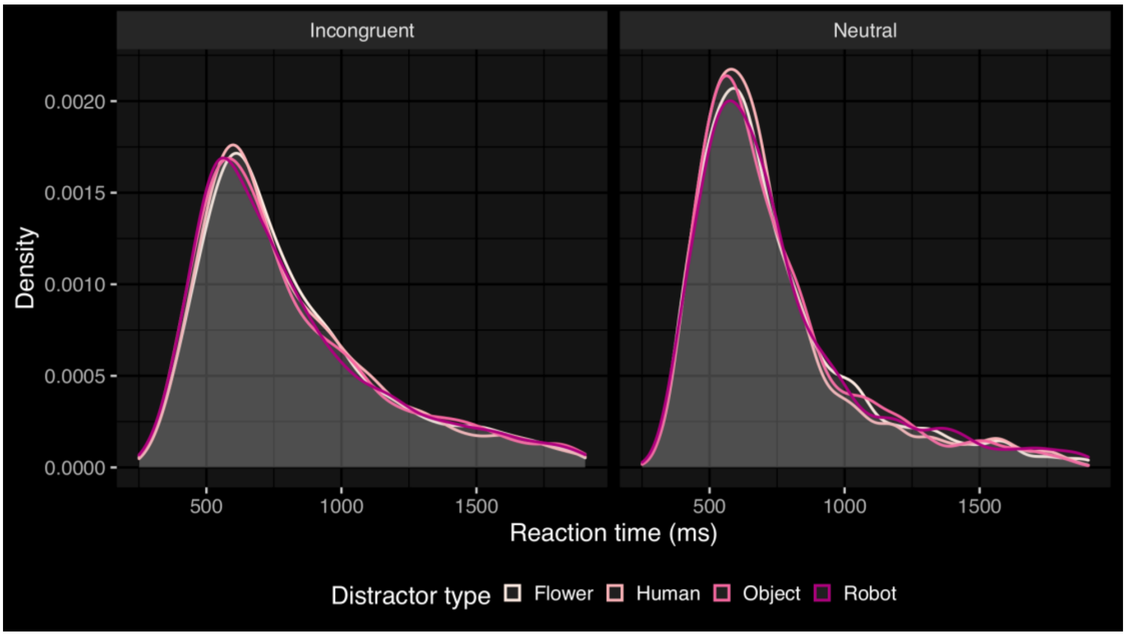
Repeating the same procedure for Experiment 2, we find that with our pre-registered standard deviation criterion we discard 1061 trials (10.84%) in total (with the participant sensitive criterion we discard 11.35% of all trials). Again, we see that the overall patterns of results remain the same across these two pre-processing methods (illustrated in tables and figures below).

T4 - Experiment 2: Number of remaining trials per experimental condition after reaction time trimming using the pre-registered standard deviation criterion. 192 trials, 51 participants.

Condition	**Total number of trials**	**Trials remaining**	**% of trials remaining**
incongruent_human	1,224	1062	86.8
incongruent_robot	1,224	1072	87.6
incongruent_object	1,224	1061	86.7
incongruent_flower	1,224	1056	86.3
neutral_human	1,224	1120	91.5
neutral_robot	1,224	1100	89.9
neutral_object	1,224	1109	90.6
neutral_flower	1,224	1101	90.0

T5 - Experiment 2: Mean reaction times (in ms) per condition using the pre-registered standard deviation criterion (Experiment 2).

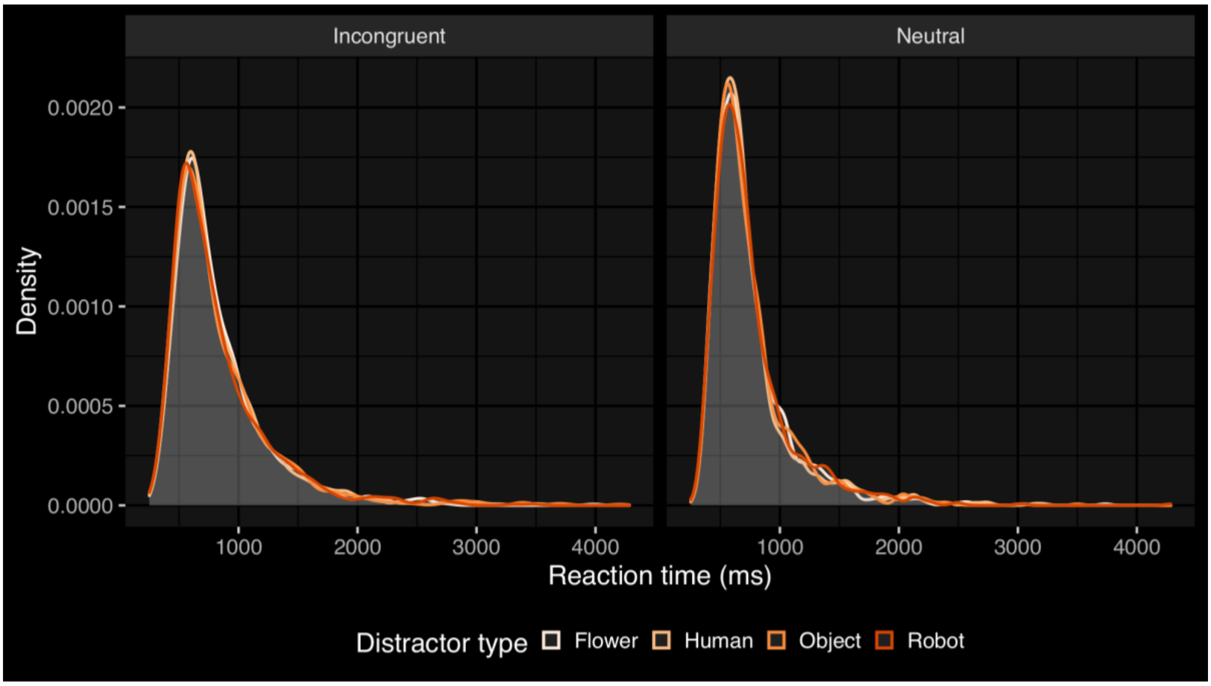
	Distractor			
	Human	Robot	Object	Flower
Incongruent target	811	808	809	816
Neutral target	723	747	730	735



F8 - Density plots for reaction times (ms) in Experiment 2 with the pre-registered standard deviation trimming criterion.

T6 - Experiment 2: Mean reaction times (in ms) per condition using the standard deviation per participant criterion.

	Distractor			
	Human	Robot	Object	Flower
Incongruent target	833	828	834	825
Neutral target	748	755	750	749



F9 - Density plots for reaction times (ms) in Experiment 2 with the per participant standard deviation trimming criterion.

Appendix D Rebuttal for "Faces do not attract more attention than non-social distractors in the Stroop task"

NB: Shared with permission from the editor. This manuscript is currently accepted pending minor revision at Collabra: Psychology (i.e. we opted for open peer review).

Editor

Summary: Three expert reviewers have provided comments on your work and find that some revisions are necessary before it would be suitable for publication. The reviewers have made suggestions that cover most of the work and many of these points are critical, while described in much more detail in the reviewer comments below, I would like to highlight a few that I found particularly important to be addressed. (1) Structure of the introduction, as detailed by Reviewer 2. (2) Reporting the number of excluded trials. (3) Further methodological details need to be included. All of the reviewers thought the figures and tables were well done.

Response: We are very grateful for the reviewers' detailed comments and suggestions and the editor's synthesis of the overarching points on how best to improve this paper. In the revised manuscript and our responses below, we detail how we have taken this feedback onboard. In line with the suggestions of the reviewers we have placed the strongest emphasis on reworking the introduction and the discussion section, providing more synthesis, critical commentary and making more explicit our study's rationale. We have detailed the exact number of excluded trials in the supplementary materials (Appendix C), as well as added the overall number of excluded trials in each experiment to their respective results sections. In order to ensure full transparency, we included the alternative reaction time pre-processing method which was recommended by one of the reviewers in the supplementary materials as well (Appendix C). Finally, we have carefully followed the reviewers' suggestions to include more details on our methods while at the same time removing some information that was flagged as redundant. We highlighted the changes to the manuscript in bold typeface.

Reviewers

Reviewer D: The authors were interested in human-robot social interaction with a focus on social motivation towards artificial agents. In two studies, they investigated the effects of distractors with varying social salience on an adaptation of a classic Stroop task. In both studies conducted, a classic Stroop effect emerged, yet there was no significant effect of salient social cues (human face distractors) capturing attention.

D.1a: The authors do a good job setting up the study's aims by providing a big picture question. However, the organization of the literature review is somewhat hard to follow. The studies reviewed contain details that do not seem necessary and distracts the reader from the main point. For example, the sample sizes each study had are unnecessary as well as step-by-step accounts of their experimental procedures. I think the introduction would benefit from succinct accounts of the main manipulations of the studies (e.g., direct vs averted gaze, open vs closed eyes) and relevant results (as the authors already do). Additionally, we do not see synthesis or commentary by the authors. It would be useful if they could provide their own interpretation of the literature and its implications.

Response: As the introduction has been criticized by all three reviewers, we have made major changes to its structure and content, now focusing less on very detailed accounts of each study’s experimental procedure and proving a bird’s eye view on the current state of the art social attentional capture research.

Thus, we have removed superfluous details such as participant numbers in the revised introduction, for example:

P.8, l. 237-239: Importantly, the authors tested the paradigm in two groups of children: typically developing boys and a group of male adolescents with Autism Spectrum Condition (ASC).

We agree that our evaluations of the literature should include more critical reflection, which we have added as well in rewriting the introduction.

P.6, l. 187-191: While the evidence on how deeply illusory faces are perceived as social is mixed, they constitute an ideal control for human facial features in social attentional capture tasks. This also raises the question how deliberate pareidolic faces, such as humanoid robots, might engage our visual attention, as these agents are capable of at least some interactions with the physical world.

The more strongly emphasized synthesis and critical commentary is especially evident in the revised discussion, for example:

P.21, l. 522-529: Many studies report effects based on very small samples (some as small as 8 participants per experiment; Ariga & Arihara, 2017; Miyazaki, Wake, Ichihara, & Wake, 2012; Sato & Kawahara, 2015), make bold statements based on modest statistical evidence ("the three-way interaction approached significance, F(2,76) = 2.46, p<.10", p. 1103, Hietanen et al., 2016) or use small sets of distractor images which are repeated across many experimental trials (Bindemann et al., 2007; Theeuwes & Van der Stigchel, 2006). Indeed, some of these problematic confounds have been highlighted and tested by Pereira and colleagues (2019; 2020).

And:

P.21, l. 539-545: While a different task was used in these studies, the authors' findings closely align with ours: faces are not reliably capturing attention and impairing the performance on an unrelated cognitive task. Interestingly, in a direct replication of Bindemann and colleagues (2007), using less well-controlled stimuli, the authors were able to replicate the effect of attentional capture by task-irrelevant faces, providing convincing evidence for systematic confounds obscuring the true picture in the existing literature.

D.1b: Refrain from using direct quotations as it takes away from the authors' original thinking. Where "centrally presented direct gaze delay[ing] attentional disengagement and recruit[ing] cognitive processing resources, and hence, processing times of the peripheral targets and the Stroop

interference are increased" is written, perhaps the authors could paraphrase the central idea.

Response: This point is well taken, and similar points have been raised by other reviewers. In the revised version of the manuscript, we have completely restructured/rewritten the introduction taking this feedback onboard.

D.1c: Details such as "The experimenter explained the procedure of the study and ensured participants understood the task" and "During this part of the study, the light in the cubicle was still switched on, and was switched off when participants started the test phase of the experiment" do not necessarily have to be in the body of the paper. If the authors would like to keep this, it would be better to move this in Supplementary Materials.

Response: We agree that this information is unnecessary and have removed it.

D.1d: It might be worth thinking about the importance of controlling for emotional valence when using human face stimuli and provide this as a motivation for using neutral stimuli, which is missing in the paper. This might be useful in the Introduction or in the General Discussion and set this up as a limitation or a note for future studies. There are several studies showing that emotional human faces (or emotional stimuli in general) have been found to capture attention faster than neutral when task-irrelevant (e.g., Theuuwes & Van der Stigchel 2006; Pessoa, McKenna, Gutierrez, & Ungerleider, 2002; Vuilleumier, 2002). This could influence the degree of social salience of social agents.

Response: Indeed, emotional valence of faces plays a crucial role in social interaction and has been repeatedly shown to influence attention differentially. Thank you for the helpful literature recommendations. We added our rationale for selecting neutral faces to the methods:

P.10, l. 299-306: The rationale behind including only neutral faces was that emotional facial cues have been shown to draw attention, especially in comparison to neutral facial expressions (Pessoa, McKenna, Gutierrez, & Ungerleider, 2002; Theeuwes & Van der Stigchel, 2006; Vuilleumier, 2002). An independent sample rated the first pool of human and robot images, resulting in a pre-selection of more neutrally perceived faces (more details can be found in the Supplementary Materials).

To illustrate the point of possibly more varied robot and object images, we show below additional stimulus examples for Reviewer D's information, which, due to copyright restrictions we cannot include in the manuscript:

[This figure had to be excluded due to copyright restrictions.]

In the revised discussion, we write:

P.23, l. 605-614: Despite our best efforts to only include neutral faces, the emotional content of the social stimuli could not be controlled to a fine-grained degree, as it was limited by the design and availability of the robots and objects that were identified through our Google search. In the emotion rating experiment, which we undertook prior to Experiment 1, the robot faces were not rated as unambiguously neutral as the human faces, even after excluding the outliers. Human faces were selected from the neutral category of the Radboud and London faces database, so these stimuli would have contained inherently less variance in perceived emotionality than the robot and object faces. However, given the scarcity of frontally oriented and high-quality robot and object faces, we chose to operate within those constraints.

D1e: Statistical analyses are sound.

D.2: (Figures, Tables, data availability) While sufficiently described, I would like to see a schematic of the Stroop task as it is always helpful to readers. The plots are beautiful.

Response: In line with this helpful suggestion, we have added a schematic representation of the Stroop task in the Methods section (p.11).

D.3: (Ethical approval) Ethical approval is present and informed consent is declared.

D.4: (Language) English is excellent. As per my comments above, the authors could be more concise when reviewing the literature, as well as describing the methods.

Response: We hope to have addressed this concern by making the changes we outline in the above responses.

Reviewer H: This manuscript featured two nearly-identical experiments designed to examine the impact of social salience during the Stroop Task. Specifically, the authors varied the social salience of distractors, including images of human faces, robot faces, objects that looked like faces, and flowers, and predicted that faces would amplify the stroop effect due to their high social salience. They found evidence for this in Experiment 1, but when they controlled for a stimulus confound in Experiment 2, they were not able to reject the null hypothesis. This investigation has some strengths. For example, comparing evaluations of human and robot faces is interesting, their analyses were appropriately simple and clear, and basic visual characteristics of the stimuli were well controlled for. The authors also took care to evaluate whether their data provided evidence in favor of the null hypothesis, which I appreciated. In general, I don't have any concerns about the methods, the data, the analyses, or their interpretation. However, I did have quite a few major concerns about more general issues that tempered my enthusiasm for the work. I'll explain these in more detail below.

H.1a: First, the majority of the Introduction is highly specific to research on gaze perception, as is the Discussion. And yet gaze direction is not examined, nor is it even important, in the current investigation. This reflected a larger issue with the Introduction, which seems to cover many topics before finally focusing on the hypothesis and aims of the investigation. It didn't feel like the gaps in the literature (as described) necessarily led to the current work and its design. Rather, it felt a bit like a literature review was forced around the current experiment. In this sense, I did not think that the article was as logically structured as it could have been.

Response: Thank you for this critical reflection – as we have written in response to the editor and the other reviewers, we have reframed both our introduction and discussion to better explain our rationale for designing the conceptual extension of the eye contact effect. We agree wholeheartedly with this criticism. The revised introduction includes a general overview of the social attentional capture focusing on whole faces (rather than gaze perception) and relates this literature back to human-robot interaction research, as this was one of the main motivators for designing this conceptual extension.

For example:

P.4, l. 101-105: Given their prioritization in our visual environment, it is unsurprising that faces have been the central focus of many visual attention studies. Collectively, these studies point towards faces ranking above objects in capturing automatic attention. Using a change blindness paradigm, Ro, Russel and Lavie (2001) found that participants detected changes in temporarily presented faces more quickly than changes in any other object.

And:

P.5, l.149-153: Hence, and as Geiger and Balas (2020) point out, robot faces, which we have presented here as a special case of intentional pareidolia, constitute a border category of face processing, and while some research exists on attentional capture by pareidolic faces, less is known about the social relevance of robot faces.

H.1b: Second, it's not clear to me that the design the authors have selected is the best one to examine their main question. In describing previous work by Conty and then Chevallier, the authors state that "the lack of difference in arousal would lead to "centrally presented direct gaze delay[ing] attentional disengagement and recruit[ing] cognitive processing resources, and hence, processing times of the peripheral targets and the Stroop interference are increased". They then go on to state that testing this claim was the goal of the current investigation. But as far as I can tell, their design does not examine arousal, nor is it confirmed that the stimuli themselves differ in the extent to which they arouse the participants. I don't have an issue with their stimulus choices per se (they're rather clever), but they don't seem to fall out of the literature reviewed, and it seems like not

manipulating gaze direction was a missed opportunity. Thus, although I have no issues with the analysis or the data, it's not clear to me that the conclusions reflect the underlying question, at least as it's framed in the introduction.

Response: This comment is very much in line with the comments of the other reviewers and editor, so we decided to reframe the introduction to clarify the rationale for our task, reduced the discussion of the follow-up experiment by Hietanen and colleagues (2016), removed the direct quote, which was criticised by another reviewer as well (see **comment D.1b**), and moved the entire section to the discussion.

In this paragraph, we were trying to establish that another conceptual extension of the eye-contact effect by Hietanen and colleagues failed to show the predicted effect: these researchers reported an effect in the opposite direction (reaction times speeding up) and credit levels of arousal in their experiment with an embodied confederate as an explanation. They describe studies with pictorial stimuli (as our studies, or the studies by Conty, Chevallier and colleagues) as low-arousal situations, in which the original effect should hold. However, we were of course unable to provide convincing evidence for a social salience effect in this version of the Stroop task. While we did not measure arousal directly, we wanted to pick up this point by Hietanen and colleagues, to continue the conversation on why a null effect could be observed in conceptual extensions of this paradigms.

P.20, l. 499-506: Hietanen and colleagues (2016) found a main effect of direct gaze speeding up the RTs of the participants, as compared to averted gaze. The authors reconcile their contradictory findings by relating them to the higher arousal produced by their stimuli: eye contact with a real person should be more engaging than pictorial representations thereof. In so-called low arousal contexts, they argue, salient social cues should recruit attentional resources and interfere with participants' performance on cognitive tasks. In our experiments, even in a context that Hietanen and colleagues (2016) describe as "low arousal", it is most probable that any social salience effect is practically equivalent to zero.

Like Chevallier and colleagues (2013), we chose not to manipulate gaze direction in this task, but rather include the neutral flower distractors (just as Chevallier and colleagues did) and vary the levels of socialness of the distractor agents. Based on the findings of Chevallier and colleagues, we expected a human distractor-dependent enhancement of the Stroop effect in the incongruent condition, compared to the flower distractor. Despite the experiments not explicitly investigating eye gaze, the gaze direction of all "social" stimuli (humans, robots, objects) was direct, towards the observer. Thus, despite taking up a smaller region in the distractor image, the direct eye gaze was controlled across social distractors, and any one of these categories should then draw more attention than the flower images. We added this point to our discussion:

P.20, l.507-519: How can our results then be explained? Of course, the stimuli we presented were more complex than those used in the original studies, so it is possible that the eye-contact effect only holds in (more) simplified contexts. The eye region in our stimulus set appeared smaller than in the original experiments, due to it taking up a smaller percentage of pixels in our distractor images. While the eye region itself was smaller, all of our social stimuli (the human, robot and object faces) depicted direct gaze and a frontally oriented face. They only varied in their potential as a social interaction partner. So, if the eye-contact effect were to hold, we should have seen a consistent difference between our most salient social stimuli with direct eye gaze (the human faces) and the neutral control condition (flowers). The fact that our data did not support this pattern is especially surprising given that past studies examining direct gaze have also used full-face stimuli in similar, cognitively demanding tasks (Burton, Bindemann, Langton, Schweinberger, & Jenkins, 2009; Conty, Russo, et al., 2010a).

H.1c: Third, critical information about the task is missing. Yes, the Stroop Task is well known and quite simple, but it isn't adequately described in the Methods, nor is any background provided about the history of the task or its mechanisms. This wouldn't be too difficult to rectify, but as it stands, it's a curious omission.

Response: We have added detailed information and figures on the number of discarded trials in the main text and supplementary materials (see also Reviewer comment **I.4**) and have added more information on the design of the Stroop task (see also Reviewer comment **I.8**).

Further to Reviewer H's request, we have added a section discussing the history of the task and its mechanisms in the introduction.

P.7, l. 208-222: Despite the above reviewed variety of paradigms which probe (social) attentional capture, the Stroop task has proven to be a particularly popular vehicle. Named after the psychologist who discovered the effect, hundreds of studies have shown that naming the ink colour of an incongruent colour word (i.e., the word "RED" presented in green) produces slower reaction times than determining the colour of a control word (the letters "XXX" presented in green). This interference effect, which highlights the fact that task-irrelevant information is processed concomitantly and automatically, has inspired a multitude of extensions, including pictorial, spatial, and social versions (MacLeod & MacDonald, 2000). For example, in the facial-emotional Stroop, participants name the ink colour of emotional, compared to neutral faces, which are overlaid with a coloured filter. Past research has shown that sad participants and participants with higher trait anger are slower to name the colour of angry versus neutral faces (Isaac et al., 2012; van Honk, Tuiten, de Haan, vann de Hout, & Stam, 2001; Van Honk et al., 2000). Thus, the Stroop task has been validated as a suitable paradigm to assess the distracting power of task-irrelevant information, such as facial cues.

H.1d: Finally, I found it hard to process the takeaway message of the manuscript. The authors found evidence against the null hypothesis in Experiment 1, but there were issues with a stimulus confound, and then in Experiment 2, there appears to be no effect of category, but the authors were at the same time not able to support a case in favor of the null--of faces not drawing more attention in the Stroop task. In other words, it's just really difficult to get a clear sense of what the study demonstrates, and thus what it's impact will be.

Response: Following comments from Reviewer I, we have revised the section on the Bayesian re-analysis of the data (including Figure 6). We hope that our interpretation of the results is now clearer: while the ROPE analysis does not offer compelling evidence in support of the null hypothesis, we can quantify our uncertainty. The 95% credible interval of the posterior distribution contains zero and overlaps to ~ 50% with our region of practical equivalence. Thus, if human faces draw more attention in the incongruent condition than the flower distractors, this effect is much smaller than expected and the evidence for it is not very strong. By providing our posteriors, other Bayesians may include them as priors and collect enough evidence to support one decision over the other. Science is cumulative, and Bayesian statistics give us an important advantage of quantifying our uncertainty, which would have not been possible if we stopped at the point of describing the null effect of the Frequentist analysis.

P.19, l.462-468: In summary, in defining our Bayesian regression model, we have increased the uncertainty of our estimates by including more random variance in the form of subject-level random effects. This increased uncertainty is expressed in Figure 5. Based on the ROPE analysis, we cannot definitively support the null hypothesis. However, considering that zero is contained in the 95% interval of credible values of the parameter's posterior distribution, the evidence for an effect is not very strong, and if real, goes in the opposite direction: -10ms [-10, 40].

H.2: (Figures, Tables, data availability) The tables and figures are nice. Well done.

H.3: (Ethical approval) This seemed adequate.

H.4: (Language) In general, yes. The quality of English was good.

Reviewer I: In this preregistered study, the question was investigated whether human faces automatically attract attention more than other types of distractors (human-like faces or non-faces) while participants solve a Stroop task. Two studies are presented, where in study 1 (N=39) a small effect seemed to favour the prediction with slightly increased Stroop effects in the presence of human faces, but a second study (N=51) that increased the number of unique distractor images failed to find differences between response times to the different distractor types.

I enjoyed reading this well-written manuscript, the theoretical background is nicely developed, the methods are sound and the statistical analyses are sophisticated. I have a few observations nevertheless that I would like the authors to consider.

I.1: The task involved the concurrent presentation of a Stroop colour-word interference test and distractors. While this setup seems to follow methods by Conty et al. (2010), the main measure involves a form of "dual distraction" - distraction from the colour-incongruent words and distraction from the faces. It would have been nice to have baseline trials in which no distractors were shown, in order to evaluate people's Stroop effect per se, without imposing a second task.

Response: Indeed, our experiments were designed as a conceptual extension of Conty and colleagues (2010a), and we followed the original procedure as closely as possible. Seeing as the Stroop effect is considered robust in the literature, we did not include another control condition without any distractor images to establish this as a ground truth. Given that we find a main effect of target in the pre-registered analysis of both experiments, we can assume that the task itself worked and, overall, induced the desired Stroop interference effect (with some variance between participants, of course).

We reemphasized this point (in addition to raising it in the abstract), by including it in the Results sections of Experiments 1 & 2:

<u>P.12, l. 355-356: This finding confirms that our modified task was still effective at inducing a Stroop interference effect.</u>

<u>P.15, l. 421-422: Again, this showed that the task worked as expected.</u>

I.2: The visual layout of these stimuli on the screen was not entirely clear to me as it is not shown in the figures, although described in the text. Was the distance between the words and the distractors different or the same as in the original study? In other words, was it perhaps easier to ignore the distractors here than in Conty et al., especially given that ignoring the distractors was indeed what participants were asked to do.

Response: We agree with the Reviewer that a visual representation of the experimental paradigm would be helpful, which we have added to the Methods section and is also visualised below:

[See Figure 8]

The distance between the distractors in our experiments and the original studies was matched as closely as possible given the difference in shape.

In trying to emulate the stimulus size, we faced the following problem: to compute size based on reported visual angle, information on the distance at which the stimulus is viewed is also necessary. This information was missing from the 2010 paper. As a workaround, we referred to the later paper by the same group, which used a similar paradigm (exchanging the averted gaze control condition for flower distractors): Chevallier et al. (2013). This allowed us to calculate the size of the distractor images using the following code in R:

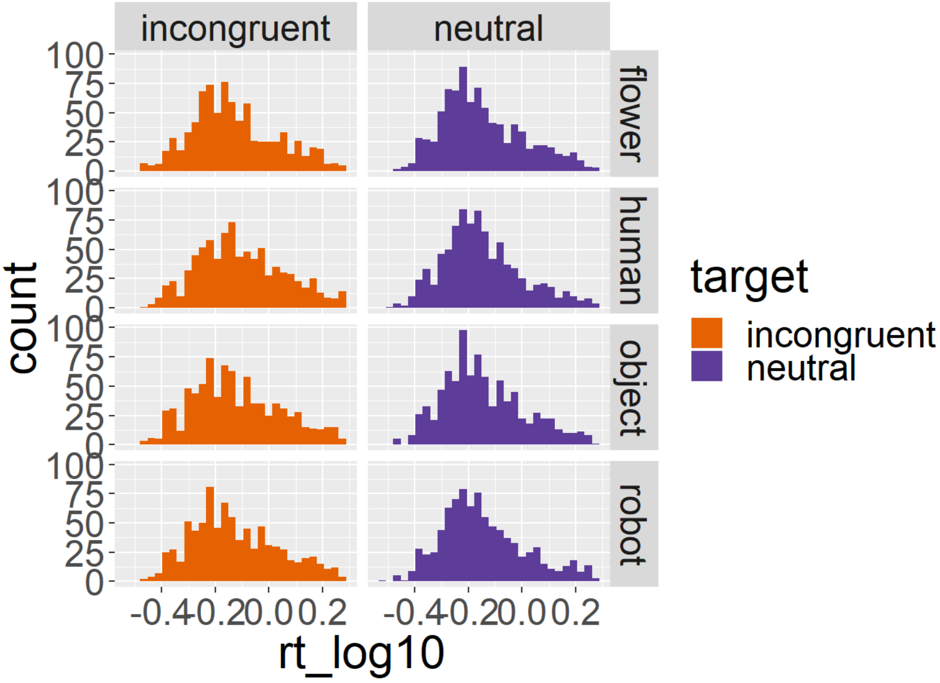
```
desiredSize <- function(visAngle, distance){
Rad = visAngle/(180/pi)
size = 2*distance*tan(Rad/2)
return(size)
}

dist=50
ang=6
desiredSize(visAngle = ang, distance = dist)
5.24
```

(Code taken from: http://stephenrho.github.io/visual-angle.html)

Thus, we can be confident that the target words and distractor images had a comparable size and were at the same distance from each other as in the original studies.

I.3: The skewed RTs (as nicely shown in Figure 4, for study 2) were analysed in raw format without further transformation (log) - have the authors tried to analyse log-transformed RTs?

Response: As the Reviewer correctly observed, here we only report the untransformed reaction times, as we did not pre-register any data transformations. However, upon initially inspecting the skew, we did try the log-transform, thus achieving an approximately normal distribution:



Log-transformed reaction time data of Experiment 1.

The log-transformation did not change the results of either of the two studies, and as a recent preprint questions the usefulness of this convention (Schramm & Rouder, 2019), we decided to focus our exploratory report on the Bayesian re-analysis of Experiment 2. In the Bayesian analysis we fit an exgaussian distribution to the data, which represents the inherent right-skew better (Baayen & Milin, 2010).

References

- Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12-28.

- Schramm, P., & Rouder, J. (2019). Are Reaction Time Transformations Really Beneficial? https://psyarxiv.com/9ksa6/.

I.4: Perhaps most critically, it seems that trials in which response times larger than 2 SD above the sample mean were excluded. This is likely too stringent since individual response times are quite variable and in fact, the most interesting trials in this task would be those in which distraction was maximal (i.e., response times are long). In order to avoid "overcleaning", I would strongly recommend either not to exclude long trials, or to use individual response time distributions - exclude trials that are 2 (or 3) SDs above each participant's own mean RT instead of the sample mean. Numbers of excluded trials and excluded trials per condition are also not reported and should be added. As a result, the remaining trials could be biased towards those that were not distracting (no matter which condition).

Response: We agree with the Reviewer that the criterion for pre-processing the reaction time data was perhaps too inflexible and did not accommodate between-participant variability. As outlined in our previous responses, we followed the procedure of the original studies and thus specified in our pre-registration: "Outlier data are defined as reaction times below 200 ms and as more than 2 standard deviations above the mean."

To investigate the concerns of the Reviewer, we used the {trimr} package, which allows the implementation of various response time trimming criteria. We used a standard deviation criterion sensitive to the participants' own means (8.9% of all trials were removed), as well as the standard deviation criterion we pre-registered (8.09% of all trials were removed). We included two tables side by side that show the means for both methods. Using the suggested standard deviation per participant criterion resulted in overall slower reaction times (Table 2), compared to the original analysis (Table 1).

[To avoid duplication, a cross-reference to above figures and tables is given below]

Appendix C

However, the pattern of results nonetheless looks similar to what we originally reported, and in the interest of adhering to our pre-registration, we decided to keep the current results section of Experiment 1 as it is, but add a table on the number of discarded trials in the supplementary materials, as well as add the percentage of the total amount of discarded trials in the main text:

P.12, l. 338-340: As a result, 606 trials (8.09%) were discarded (a detailed breakdown of the trial number per condition can be found in the Supplementary Materials).

Repeating the same procedure for Experiment 2, we find that with our pre-registered standard deviation criterion we discard 1061 trials (10.84%) in total (with the participant sensitive criterion we discard 11.35% of all trials). We added the information on discarded trials in the main text:

P,15, l. 410-412: With this reaction time trimming criterion, we discarded 1061 trials (10.84%). A detailed breakdown of the number of trials remaining per condition can be found in the Supplementary Materials.

For comparison, we list Table 5 and 6 with the mean reaction times for Experiment 2 using our reported method and the method recommended by the Reviewer:

[To avoid duplication, a cross-reference to above figures and tables is given below]

Appendix C

Again, we see that the overall patterns of results remain the same across these two pre-processing methods.

References

Grange, J. A. (2015). trimr: An implementation of common response time trimming methods. R package version 1.0. 1

I.5: The Bayesian analyses are sufficiently esoteric for me that I require more clarification here.

Response: We agree with the Reviewer that the report of the exploratory Bayesian analysis in its original form was not as clear as it could (or should) have been. We hope we have addressed this concern sufficiently in the following responses.

I.5a: Page 13 states "Given the results of Study 2, we explored the extent to which our data provided compelling evidence for the null hypothesis (no difference in reaction times in the incongruent and neutral conditions when human faces are presented)". This implies to me that the null hypothesis would predict no Stroop effect when the human faces were presented. I believe this is not what the authors meant, but instead that the size of the Stroop effect would not differ between distractor conditions. Is this the case? If so, this needs to be changed in the text.

Response: We thank the reviewer for pointing out this mistake. We have amended the text accordingly:

P.16, l. 428-433: Given the results of Experiment 2, we explored the extent to which our data provided compelling evidence for the null hypothesis (no enhanced Stroop effect when human faces are presented compared to the control flower condition) by using a Bayesian regression modelling approach ({brms} package in R and Stan (Version 2.9.0), Bürkner, 2017), as the null cannot be confirmed with Frequentist statistics.

I.5b: How are R-hat values of 1.00 for each of the tested parameters in Table 3 to be understood?

Response: The R-hat value provides information on how well the algorithm could estimate the posterior distribution of each parameter. Since we already provided information in the main text on the convergence of the model, this column has been removed to avoid redundancy.

I.5c: The ROPE outcomes do not support the presence of a Stroop effect at all, if I understood this correctly. The size of the general Stroop effect was sufficiently large, in both studies, based on the conventional outcomes (study 1: F(1, 38)= 39.24, p<.001, ηG2= .03; study 2: F(1, 50)=70.31, p<.001, ηG2=.06). Can the authors comment more directly on this discrepancy? And if the outcomes of the Bayesian analysis are taken seriously, what are the consequences for the rest of the paper? For example, page 15 in the discussion states "While we again observed the predicted Stroop effect" - did you? The different outcomes need to be reconciled better, in my opinion.

Response: We thank the Reviewer for raising this issue. Determining the limits of the ROPE is a somewhat controversial issue in the literature (Kruschke, 2018; Kelter, 2020). Instead of relying on the automated procedure we opted for initially via the {BayesfactoR} package's rope_range() function, in the revised manuscripts, we choose the limits based on half of what is considered a small effect. The rope_range() function returned a range that is considered a large effect and because of this, the robust Stroop effect was classified as "undecided". Based on Experiment 1 (Δ 47ms) and the findings of Conty et al. (Δ 34ms) and Chevallier et al. (Δ 41ms), we set the ROPE limits to [-.017, .017] and observe a plot that is easier to reconcile with the Frequentist analysis.

P.18, l. 451-457: In determining the ROPE range, we set the limits following the procedure based on half of what we consider a small effect (Kruschke, 2018). A small effect in our first experiment was an average difference of 47ms between the incongruent social and incongruent control distractor, compared to a difference in 34ms in Conty and colleagues' task and 41ms in Chevallier and colleagues' version (2010, 2013). Choosing the most conservative small effect, we set the ROPE limits to [-.017, .017].

I5.d: Figure 5 is unclear to me. What is zero on the x-axis - this can't be "reaction time (s)" ? Also, going back to point 4 a) is this testing the presence/absence of any Stroop effect or the slow-down of RTs (i.e., bigger Stroop effect) for human faces compared to the other conditions?

Response: We included an updated version of Figure 5, now with a corrected x-axis label. Thank you for pointing out this issue. The graph depicts all parameters estimated in the Bayesian regression model. So, for instance, the estimated effect of the incongruent target on the reaction time (the Stroop interference effect), for which H0 is rejected. We also see the parameter estimates for the different distractor types, as well as the interactions. The effect of interest (as outlined in the figure description in the manuscript) is the incongruent target with the human distractor type. This effect is now shaded in yellow, and we do not have a clear decision on H0 based on the ROPE analysis. However, as we have written in our updated Bayesian analysis section, the estimated effect is small and likely not very strong (if present at all). Moreover, it is smaller than the smallest effect we consider interesting (based on our previous experiment and the literature), and in the 95% CI zero is contained as a likely value.

P.19, l. 462-468: In summary, in defining our Bayesian regression model, we have increased the uncertainty of our estimates by including more random variance in the form of subject-level random effects. This increased uncertainty is expressed in Figure 5. Based on the ROPE analysis, we cannot definitively support the null hypothesis. However, considering that zero is contained in the 95% interval of credible values of the parameter’s posterior distribution, and more than 50% of its values are practically equivalent with zero, the evidence for an effect is not very strong and even goes in the opposite direction: 10ms [-.01, .04].

[See Figure 12]

I.6: Using mirror-images also in study 2 arguably may not have created unique distractors. A mirror image could act as a particularly strong distractor, as it would appear familiar but not identical. This could be considered in the limitations section.

Response: We agree that this point should be raised in the limitations section and have added it:

P.23, l. 593-604: A further stimulus-based limitation was that in Experiment 1, distractors were not controlled by their mirror and presented twice. Thus, the repeat presentation could have led to a particularly memorable stimulus set. In Experiment 2, the unique distractors in the incongruent condition were controlled by their mirror images. Of course, on the other hand, the repeat presentation of distractor images is common practice in the social attentional capture literature (for example, a set of four unique human and pareidolic face images used for an experiment consisting of 450 trials, Ariga & Arihara, 2017). Takahashi and colleagues (2013) used stimuli with three unique identities over many trials, and only four unique stimuli in another study (Takahashi & Watanabe, 2015). Theeuwes et al. (2006) presented 12 unique distractor images across 96 trials. To put it differently, based on the conventions of the social attentional biasing literature, it is unlikely that we did not observe the expected effect due to the number of unique distractor images we presented.

I.7: The decision to move the stimulus rating into a supplement abbreviated the rating outcomes presented in the paper, but I would still have liked to see some details. In fact, the supplement also does not state what exactly was being judged regarding these stimuli. The paper states on page 11, “mind perception of different kinds of agents” – what does this mean and what was the actual outcome of the ratings? Is it relevant or irrelevant for this paper?

Response: We have tried to clarify our rationale for the ratings of our distractor images in Experiment 2. As we have written, we wanted to establish that the 4 different categories were perceived differently with regard to “having a mind”, which we implicitly equated with the agent’s potential for socialness. The two items (and their descriptions), which we called “agency” (ability to plan and act) and “experience” (the ability to sense and feel), were derived from Gray, Gray & Wegner (2007).

P.14, l. 402-407: We used mind perception as a socialness proxy to distinguish between the control condition (flowers), inanimate (robot and pareidolic faces) and agents with a mind (humans). The analysis of the ratings confirmed that the stimulus categories were perceived differently: the human images received the highest agency and experience ratings. A detailed report of the stimulus ratings can be found in the Supplementary Materials.

I.8: Some missing details on the Stroop task itself included the number/ratio of congruent and incongruent trials, and any restrictions regarding the switch between the two (e.g., no more than 2 incongruent trials after each other etc.)

Response: We have added this information in the methods section:

P.11, l. 327-329: There were equal numbers of incongruent and neutral Stroop trials, and no restrictions regarding the switch between incongruent and neutral trials were put in place (as they were presented randomly). The target word and distractor image pairs were fixed.

I.9: Since several participants were excluded, I wonder whether these criteria were too stringent. At least the method of excluding participants should be detailed. For example, excluding participants with ASD diagnoses – how was this done?

Response: We have added this information.

The rationale for specifying these exclusion criteria was that Chevallier and colleagues (2013) found diverging results for the ASD participant group in their sample, and we wanted to ensure that all participants were equally naïve towards robots (as the initial goal was to establish this as a robust measure for social motivation, and then in future experiment integrate this task following prolonged human-robot interaction. We were curious about seeing any potential differences between a robot-naïve group of participants and a group that has encountered humanoid robots on this task).

P.9, l. 284-289: We recruited 50 participants, however, based on our pre-registered exclusion criteria (diagnosis of ASD and having had a previous interaction with a robot) we excluded 9 participants. Two additional participants had insufficient English language skills, and thus the total number of exclusions was 11. The pre-registered exclusions were made based on participant answers on the experiment questionnaires' self-report items (for example: "Do you have a diagnosis of Autism Spectrum Disorder?" and "Have you interacted with a robot before?"). The other exclusions had to be made in addition, based on difficulties participants had with the task. We report a final sample size of N=39.

I.10: (Figures, tables, data availability) Very nice use and high quality of Figures.

I.11: (Ethical approval) Ethical approval was obtained from the University of Glasgow ethics review board (300170224).

I.12: (Language) English is appropriate.

References

Aaltonen, I., Arvola, A., Heikkilä, P., & Lammi, H. (2017). Hello Pepper, May I Tickle You?: Children's and Adults' Responses to an Entertainment Robot at a Shopping Mall. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 53–54. https://doi.org/10.1145/3029798.3038362

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F. A., … Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*, 63. https://doi.org/10.12688/wellcomeopenres.15191.1

Anderson, B. A. (2016). Social reward shapes attentional biases. *Cognitive Neuroscience*, *7*(1–4), 30–36. https://doi.org/10.1080/17588928.2015.1047823

Ariga, A., & Arihara, K. (2017). Visual attention is captured by task-irrelevant faces, but not by pareidolia faces. *2017 9th International Conference on Knowledge and Smart Technology: Crunching Information of Everything, KST 2017*, *1*, 266–269. https://doi.org/10.1109/KST.2017.7886105

Arslan, R. C., Walther, M. P., & Tata, C. S. (2019). formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, 1–37. https://doi.org/10.3758/s13428-019-01236-y

Asimov, I. (1941). Liar! *Astounding Science Fiction*.

Avants, B., Epstein, C., Grossman, M., & Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Balconi, M., & Molteni, E. (2016). Past and future of near-infrared spectroscopy in studies of emotion and social neuroscience. *Journal of Cognitive Psychology*, *28*(2), 129–146. https://doi.org/10.1080/20445911.2015.1102919

Bartneck, C., Kuli, D., & Croft, E. (2009). *Measurement Instruments for the Anthropomorphism , Animacy , Likeability , Perceived Intelligence , and Perceived Safety of Robots*. 71–81. https://doi.org/10.1007/s12369-008-0001-3

Baum, F. L. (1900). *The wonderful wizard of Oz*. G.M. Hill Co.

Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., & Belpaeme, T. (2016). From characterising three years of HRI to methodology and reporting recommendations. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 391–398. https://doi.org/10.1109/HRI.2016.7451777

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Belpaeme, T. (2020). Advice to new human-robot interaction researchers. In *Human-Robot Interaction: Evaluation Methods and Their Standardization* (Vol. 12). Springer Nature.

Bendahan, N., Neal, O., & Ross-White, A. (2019). Relationship Between Near-Infrared Spectroscopy-Derived Cerebral Oxygenation and Delirium in Critically Ill Patients: A Systematic Review. *Journal of Intensive Care Medicine*, *34*(6), 514–520.

Berniere, F., Reznick, S., & Rosenthal, R. (1988). *Synchrony, Pseudosynchrony and Dissynchrony Measuring the Entrainment Process in Mother Infant Interactions.*

Bethel, C. L., & Murphy, R. R. (2010). Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics*, 2(4), 347–359. https://doi.org/10.1007/s12369-010-0064-9

Bhandari, R., Kirilina, E., Caan, M., Suttrup, J., De Sanctis, T., De Angelis, L., Keysers, C., & Gazzola, V. (2020). Does higher sampling rate (multiband + SENSE) improve group statistics—An example from social neuroscience block design at 3T. *NeuroImage*, *213*(September 2019), 116731. https://doi.org/10.1016/j.neuroimage.2020.116731

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in cognitive sciences*, *23*(5), 365-368.

Bindemann, M., Burton, A. M., Langton, S. R. H., Schweinberger, S. R., & Doherty, M. J. (2007). The control of attention to faces. *Journal of Vision*, *7*(10), 1–8. https://doi.org/10.1167/7.10.15

Birhane, A. (2017). Descartes was wrong: ‘A person is a person through other persons’. *Aeon*. https://aeon.co/ideas/descartes-was-wrong-a-person-is-a-person-through-other-persons

Bjornsdottir, R. T., & Rule, N. O. (2017). The Visibility of Social Class From Facial Cues. *Journal of Personality and Social Psychology*, *113*(4), 530–546. https://doi.org/10.1037/pspa0000091.supp

Bolis, D., & Schilbach, L. (2018). 'I Interact Therefore I Am': The Self as a Historical Product of Dialectical Attunement. *Topoi*, *0*(0), 1–14. https://doi.org/10.1007/s11245-018-9574-0

Bradley, M. M., & Lang, P. P. J. (1999). Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. *Psychology*, *Technical*(C-1), 0. https://doi.org/10.1109/MIC.2008.114

Bratt, S. (2017). Toward an open data repository and meta-analysis of cognitive data using fNIRS studies of emotion. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10284*, 449–467. https://doi.org/10.1007/978-3-319-58628-1_34

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, *42*, 167–175. https://doi.org/10.1016/S0921-8890(02)00373-1

Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. *Intelligent Robots and Systems*, 858–863. https://doi.org/10.1109/IROS.1999.812787

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *Neuroimage*, 16(2), S497.

Broadbent, E. (2017). Interactions With Robots: The Truths We Reveal About Ourselves. *Annu. Rev. Psychol*, *68*(9), 1–926. https://doi.org/10.1146/annurev-psych-010416-043958

Brough, J., Henschel, A., Rabagliati, H., Harris, L., Cross, E. S., & Branigan, H. P. (2020, October 9). The influence of animacy on perspective-taking and word order during language production. Retrieved from osf.io/a5fby.

Bubic, A., Susac, A., & Palmovic, M. (2014). Keeping our eyes on the eyes: The case of Arcimboldo. *Perception*, *43*(5), 465–468. https://doi.org/10.1068/p7671

Bulgarelli, C., Blasi, A., Arridge, S., Powell, S., de Klerk, C. C. J. M., Southgate, V., Brigadoi, S., Penny, W., Tak, S., & Hamilton, A. (2018). Dynamic causal modelling on infant fNIRS data: A validation study on a simultaneously recorded fNIRS-fMRI dataset. *NeuroImage*, *175*(April), 413–424. https://doi.org/10.1016/j.neuroimage.2018.04.022

Bürkner, P. (2016). *Package 'brms'*. https://github.com/paul-buerkner/brms

Burra, N., Framorando, D., & Pegna, A. J. (2018). Early and late cortical responses to directly gazing faces are task dependent. *Cognitive, Affective and Behavioral Neuroscience*, *18*(4), 796–809. https://doi.org/10.3758/s13415-018-0605-5

Burton, A. M., Bindemann, M., Langton, S. R. H., Schweinberger, S. R., & Jenkins, R. (2009). Gaze Perception Requires Focused Attention: Evidence From an Interference Task. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 108–118. https://doi.org/10.1037/0096-1523.35.1.108

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365-376.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015.

Nature Human Behaviour, 2(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Cameron, D., Fernando, S., Collins, E. C., Millings, A., Szollosy, M., Moore, R., Sharkey, A., & Prescott, T. (2017). You made him be alive: Children's perceptions of animacy in a humanoid robot. *Conference on Biomimetic and Biohybrid Systems*, 73–85.

Campa, R. (2016). The Rise of Social Robots: A Review of the Recent Literature. *Journal of Evolution and Technology*, *26*(1), 106–113.

Cangelosi, A., & Schlesinger, M. (2018). From Babies to Robots: The Contribution of Developmental Robotics to Developmental Psychology. *Child Development Perspectives*, *12*(3), 6.

Canning, C., & Scheutz, M. (2013). Functional Near-Infrared Spectroscopy in Human-Robot Interaction. *Journal of Human-Robot Interaction*, 2(3), 62–84. https://doi.org/10.5898/JHRI.2.3.Canning

Carp, J. (2012). *The secret lives of experiments: Methods reporting in the fMRI literature*. 12.

Cervera, E. (2019). Try to Start It! The Challenge of Reusing Code in Robotics Research. *IEEE ROBOTICS AND AUTOMATION LETTERS*, *4*(1), 8.

Chambers, C., Dienes, Z., McIntosh, R., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex*, *66*(1–2). https://doi.org/10.1016/j.cortex.2015.03.022

Chaminade, T., & Cheng, G. (2009). Social cognitive neuroscience and humanoid robotics. *Journal of Physiology Paris*, *103*(3–5), 286–295. https://doi.org/10.1016/j.jphysparis.2009.08.011

Chaminade, T., & Okka, M. M. (2013). Comparing the effect of humanoid and human face for the spatial orientation of attention. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00012

Chance, B., Zhuang, Z., Unah, C., Alter, C., & Lipton, L. (1993). Cognition-activated low-frequency modulation of light absorption in human brain. *Proc. Natl. Acad. Sci. USA*, 5.

Chaudhury, B., Hortensius, R., Hoffmann, M., & Cross, E. S. (2020). *Tracking Human Interactions with a Commercially-available Robot over Multiple Days: A Tutorial* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fd3h2

Chesher, C., & Andreallo, F. (2020). Robotic Faciality: The Philosophy, Science and Art of Robot Faces. *International Journal of Social Robotics*, *January*. https://doi.org/10.1007/s12369-020-00623-2

Chevalier, P., Kompatsiari, K., Ciardo, F., & Wykowska, A. (2020). Examining joint attention with the use of humanoid robots-A new approach to study fundamental mechanisms of social cognition. *Psychonomic Bulletin & Review*, *27*(2), 217–236. https://doi.org/10.3758/s13423-019-01689-4

Chevallier, C., Huguet, P., Happé, F., George, N., & Conty, L. (2013). Salient social cues are prioritized in autism spectrum disorders despite overall decrease in social attention. *Journal of Autism and Developmental Disorders*, *43*(7), 1642–1651. https://doi.org/10.1007/s10803-012-1710-x

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in Cognitive Sciences*, *16*(4), 231–238. https://doi.org/10.1016/j.tics.2012.02.007

Chevallier, C., Tonge, N., Safra, L., Kahn, D., Kohls, G., Miller, J., & Schultz, R. T. (2016). Measuring social motivation using signal detection and reward responsiveness. *PLoS ONE*, *11*(12), 1–14. https://doi.org/10.1371/journal.pone.0167024

Ciardo, F., Beyer, F., De Tommaso, D., & Wykowska, A. (2020). Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition*, *194*, 104109.

Clabaugh, C., & Matarić, M. (2018). Robots for the people, by the people: Personalizing human-machine interaction. *SCIENCE ROBOTICS*, 3.

Clausner, T., Dalal, S. S., & Crespo-García, M. (2017). Photogrammetry-based head digitization for rapid and accurate localization of EEG electrodes and MEG fiducial markers using a single digital SLR camera. *Frontiers in Neuroscience*, *11*(MAY), 1–12. https://doi.org/10.3389/fnins.2017.00264

Collins, E. C. (2019). Drawing parallels in human–other interactions: a trans-disciplinary approach to developing human–robot interaction methodologies. *Philosophical Transactions of the Royal Society B*, *374*(1771), 20180433.

Collins, E. C., Prescott, T. J., Mitchinson, B., & Conran, S. (2015). MIRO: A versatile biomimetic edutainment robot. *Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology - ACE '15*, 1–4. https://doi.org/10.1145/2832932.2832978

Collodi, C. (1883). *Le Avventure di Pinocchio: Storia di un Buratino*. Felice Paggi Libraio-Editore.

Conty, L., Gimmig, D., Belletier, C., George, N., & Huguet, P. (2010a). The cost of being watched: Stroop interference increases under concomitant eye contact. *Cognition*, *115*(1), 133–139. https://doi.org/10.1016/j.cognition.2009.12.005

Conty, L., Russo, M., Loehr, V., Hugueville, L., Barbu, S., Huguet, P., Tijus, C., & George, N. (2010b). The mere perception of eye contact increases arousal during a word-spelling task. *Social Neuroscience*, *5*(2), 171–186. https://doi.org/10.1080/17470910903227507

Cox, R., & Hyde, J. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, *10*(4–5), 171–178.

Cross, E.S., Hortensius R, Wykowska A. (2019a) From social brains to social robots: applying neurocognitive insights to human-robot interaction. *Phil. Trans. R. Soc. B*. 374: 20180024. http://dx.doi.org/10.1098/rstb.2018.0024

Cross, E. S., Liepelt, R., Antonia, A. F., Parkinson, J., Ramsey, R., Stadler, W., & Prinz, W. (2012). Robotic movement preferentially engages the action observation network. *Human Brain Mapping*, *33*(9), 2238–2254. https://doi.org/10.1002/hbm.21361

Cross, E. S., & Ramsey, R. Mind meets machine: Towards a cognitive science of human-machine interactions. Under review.

Cross, E. S., Ramsey, R., Liepelt, R., Prinz, W., & Hamilton, A. F. D. C. (2016). The shaping of social perception by stimulus and knowledge cues to human animacy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150075–20150075. https://doi.org/10.1098/rstb.2015.0075

Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B., & Hortensius, R. (2019b). A neurocognitive investigation of the impact of socialising with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B*. https://doi.org/10.1098/rstb.2018.0034

Cross, L., Wilson, A. D., & Golonka, S. (2016). How moving together brings us together: When coordinated rhythmic movement affects cooperation. *Frontiers in Psychology*, *7*(DEC), 1–13. https://doi.org/10.3389/fpsyg.2016.01983

Crouzet, S. M., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, *10*(2010), 1–17. https://doi.org/10.1167/10.4.16.Introduction

Cui, X., Bray, S., Bryant, D. M., Glover, G. H., & Reiss, A. L. (2011). A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *NeuroImage*, *54*(4), 2808–2821. https://doi.org/10.1016/j.neuroimage.2010.10.069

Darda, K. M., Butler, E. E., & Ramsey, R. (2018). Functional Specificity and Sex Differences in the Neural Circuits Supporting the Inhibition of Automatic Imitation. *Journal of Cognitive Neuroscience*, *30*(6), 914–933. https://doi.org/10.1162/jocn_a_01261

Darling, K. (2015). 'Who's Johnny?'Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. *Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy (March 23, 2015). ROBOT ETHICS*, 2.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1480), 679–704. https://doi.org/10.1098/rstb.2006.2004

DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, *17*(6–7), 790–811. https://doi.org/10.1080/13506280902793843

DeBruine, L., & Jones, B. (2017). *Face Research Lab London Set*. https://figshare.com/articles/Face_Research_Lab_London_Set/5047666

Dereshev, D., Kirk, D., Matsumura, K., & Maeda, T. (2019). Long-Term Value of Social Robots through the Eyes of Expert Users. *Proceedings of the 2019*

CHI Conference on Human Factors in Computing Systems - CHI '19, 1–12. https://doi.org/10.1145/3290605.3300896

DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002). All robots are not created equal: The design and perception of humanoid robot heads. *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS*, 321–326. https://doi.org/10.1145/778712.778756

DiSalvo, C. F., & Gemperle, F. (2003). From Seduction to Fulfillment: The Use of Anthropomorphic Form in Design. *Proceedings of the International Conference on Designing Pleasurable Products and Interfaces*, 67–72.

Dragicevic, P. (2016). Fair Statistical Communication in HCI. In J. Robertson & M. Kaptein (Eds.), *Modern Statistical Methods for HCI* (pp. 291–330). Springer International Publishing. https://doi.org/10.1007/978-3-319-26633-6_13

Duffy, B R, Rooney, C. F. B., Hare, G. M. P. O., & Donoghue, R. P. S. O. (1999). What is a Social Robot ? *Computer*, 1–3.

Duffy, Brian R. (2000). The social robot paradox. *PhD Thesis, November*, 288. https://doi.org/10.1.1.79.3188

Duffy, Brian R. (2004). The social robot paradox. *Position Paper for the Workshop Sociality with Machines. Shaping Relationsships with Machines*.

Duffy, Brian R. (2006). Fundamental Issues in Social Robotics. *International Review of Information Ethics (IRIE)*, *6*(March 2003), 31–36. https://doi.org/10.1177/1059712316668238

Duffy, Brian R, & Joue, G. (2005). The Paradox of Social Robotics: A Discussion. *AAAI Fall 2005 Symposium on Machine Ethics, Hyatt Regency*.

Dworkin, J. D., Linn, K., Teich, E., Zurn, P., Shinohara, R., & Bassett, D. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, 14.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When We Need A Human: Motivational Determinants of Anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Eriksson, J., Matarić, M. J., & Winstein, C. J. (2005). Hands-off assistive robotics for post-stroke arm rehabilitation. *Proceedings of the 2005 IEEE 9th International Conference on Rehabilitation Robotics*, *2005*, 21–24. https://doi.org/10.1109/ICORR.2005.1501042

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Eyssel, F. (2017). An experimental psychological perspective on social robotics. *Robotics and Autonomous Systems*, *87*, 363–371. https://doi.org/10.1016/j.robot.2016.08.029

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, *51*(4), 724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

Fasola, J., & Matarić, M. J. (2012). Using socially assistive human-robot interaction to motivate physical exercise for older adults. *Proceedings of* *the IEEE*, *100*(8), 2512–2526. https://doi.org/10.1109/JPROC.2012.2200539

Feil-Seifer, D., & Matarić, M. J. (2011). Socially assistive robotics. *Robotics & Automation Magazine, IEEE*, *18*(1), 24–31. https://doi.org/10.1109/ICORR.2005.1501143

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage*, *63*(2), 921–935.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781.

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, *37*(4), 571–583. https://doi.org/10.1068/p5705

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2011). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *54*, 313–327.

Framorando, D., George, N., Kerzel, D., & Burra, N. (2016). Straight gaze facilitates face processing but does not cause involuntary attentional capture. *Visual Cognition*, *24*(7–8), 381–391. https://doi.org/10.1080/13506285.2017.1285840

Friston, K. J., Mechelli, A., Turner, R., & Price, C. J. (2000). Nonlinear Responses in fMRI: The Balloon Model, Volterra Kernels, and Other Hemodynamics. *NeuroImage*, *12*(4), 466–477.

Frumer, Y. (2020, May 21). The Short, Strange Life of the First Friendly Robot. *IEEE SPECTRUM*. https://spectrum.ieee.org/robotics/humanoids/the-short-strange-life-of-the-first-friendly-robot

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in*

Psychological Science, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of ‘theory of mind’. *Trends in Cognitive Sciences*, *7*(2), 77–83. https://doi.org/10.1016/S1364-6613(02)00025-6

Gardner, T., Goulden, N., & Cross, E. S. (2015). Dynamic modulation of the action observation network by movement familiarity. *Journal of Neuroscience*, *35*(4), 1561-1572.

Garland, A. (2014). *Ex Machina*. A24.

Gates, B. (2008). A robot in every home. *Scientific American*, *296*(1), 58–65.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage*, *35*(4), 1674–1684. https://doi.org/10.1016/j.neuroimage.2007.02.003

Geiger, A., & Balas, B. (2020). Not quite human, not quite machine: Electrophysiological responses to robot faces. *bioRxiv*. https://doi.org/10.1101/2020.06.11.145979

Georgiou, T., Singh, K., Baillie, L., & Broz, F. (2020). Small Robots With Big Tasks: A Proof of Concept Implementation Using a MiRo for Fall Alert. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 206–208. https://doi.org/10.1145/3371382.3378331

Ghiglino, D., De Tommaso, D., Willemse, C., Marchesi, S., & Wykowska, A. (2020). *Can I get your (robot) attention? Human sensitivity to subtle hints of human-likeness in a humanoid robot’s behavior* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/kfy4g

Gilder, T. S. E., & Heerey, E. A. (2018). The Role of Experimenter Belief in Social Priming. *Psychological Science*, 095679761773712–095679761773712. https://doi.org/10.1177/0956797617737128

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Gobbini, M. I., Gentili, C., Ricciardi, E., Bellucci, C., Salvini, P., Laschi, C., Guazzelli, M., & Pietrini, P. (2011). Distinct Neural Systems Involved in Agency and Animacy Detection. *Journal of Cognitive Neuroscience*, *23*(8), 1911–1920. https://doi.org/10.1162/jocn.2010.21574

Goodrich, M. A., & Schultz, A. C. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, *1*(3), 203–275. https://doi.org/10.1561/1100000005

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., ... & Hampton, J. M. (2017). Precision functional mapping of individual human brains. *Neuron*, *95*(4), 791-807.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Gowen, E. (2016). *Believe it or not: Moving non-biological stimuli believed to have human origin can be represented as human movement*. 8.

Grange, J. A. (2015). *trimr: An implementation of common response time trimming methods* (1.0.1) [Computer software]. https://cran.r-project.org/web/packages/trimr/

Greene, D. J., Koller, J. M., Hampton, J. M., Wesevich, V., Van, A. N., Nguyen, A. L., Hoyt, C. R., McIntyre, L., Earl, E. A., Klein, R. L., Shimony, J. S., Petersen, S. E., Schlaggar, B. L., Fair, D. A., & Dosenbach, N. U. F. (2018). Behavioral interventions for reducing head motion during MRI scans in children. *NeuroImage*, *171*, 234–245. https://doi.org/10.1016/j.neuroimage.2018.01.023

Greve, D. N., & Fischl, B. (2009). *Accurate and robust brain image alignment using boundary-based registration*. 10.

Guido, G., Pichierri, M., Pino, G., & Nataraajan, R. (2019). Effects of face images and face pareidolia on consumers' responses to print advertising: An empirical investigation. *Journal of Advertising Research*, *59*(2), 219–231. https://doi.org/10.2501/JAR-2018-030

Halloy, J., Sempo, G., Caprari, G., Rivault, C., Asadpour, M., T??che, F., Sa??d, I., Durier, V., Canonge, S., A., J. M., Detrain, C., Correll, N., Martinoli, A., Mondada, F., Siegwart, R., & Deneubourg, J. L. (2007). Social integration of robots into groups of cockroaches to control self-organized choices. *Science*, *318*(5853), 1155–1158. https://doi.org/10.1126/science.1144259

Hannibal, G., & Weiss, A. (2020). Envisioning social robotics: Current challenges and new interdisciplinary methodologies. *Interaction Studies*, *21*(1), 1–6.

Hayward, D. A., Voorhies, W., Morris, J. L., Capozzi, F., & Ristic, J. (2017). Staring Reality in the Face: A Comparison of Social Attention Across Laboratory and Real World Measures Suggests Little Common Ground. *Canadian Journal of Experimental Psychology*, *71*(3), 212–225. https://doi.org/10.1037/cep0000117

Heider, F., & Simmel, F. (1944). *An Experimental Study of Apparent Behavior*. *57*(2), 243–259.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *22*(2–3), 61–83.

Henschel, A. (2019). My Robotic Research Assistant [Blog]. *University of Glasgow PGR Blog*. https://uofgpgrblog.com/pgrblog/2019/3/12/my-robotic-research-assistant

Henschel, A., & Cross, E. S. (2020a). No evidence for enhanced likeability and social motivation towards robots after synchrony experience. *Interaction Studies*, *21*(1), 7-23. https://doi.org/10.1075/is.19004.hen

Henschel, A., Hortensius, R., & Cross, E. S. (2020b). Social Cognition in the Age of Human – Robot Interaction. *Trends in Neurosciences*, 43(6), 1–12. https://doi.org/10.1016/j.tins.2020.03.013

Henschel, A., Kent, M., Leisten, L. M., Seelemeyer, H., Timmerman, R., Hortensius, R., & Cross, E. S. (2020c). *A manual to digitising fNIRS probes with photogrammetry as described in Clausner et. Al. (2017)*. Zenodo. http://doi.org/10.5281/zenodo.4146985

Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? A review and perspective. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-020-01715-w

Hietanen, J. K., Myllyneva, A., Helminen, T., & Lyyra, P. (2016). The Effects of Genuine Eye Contact on Visuospatial and Selective Attention. *Journal of*

Experimental Psychology, *41*(6), 573–575. https://doi.org/10.1007/s001080050562

Hinz, A.-N., Ciardo, F., & Wykowska, A. (2019). Individual differences in attitude toward robots predict behavior in human-robot interaction. *International Conference on Social Robotics*, 64–73.

Hockstein, N. G., Gourin, C. G., Faust, R. A., & Terris, D. J. (2007). A history of robots: From science fiction to surgical robotics. *Journal of Robotic Surgery*, *1*(2), 113–118. https://doi.org/10.1007/s11701-007-0021-2

Hoffman, G. (2019, May 1). Anki, Jibo, and Kuri: What We Can Learn from Social Robots That Didn't Make It. *IEEE SPECTRUM*. https://spectrum.ieee.org/automaton/robotics/home-robots/anki-jibo-and-kuri-what-we-can-learn-from-social-robotics-failures

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). *Tenzing: Documenting contributorship using CRediT* [Preprint]. MetaArXiv. https://doi.org/10.31222/osf.io/b6ywe

Homölle, S., & Oostenveld, R. (2019). Using a structured-light 3D scanner to improve EEG source modeling with more accurate electrode positions. *Journal of Neuroscience Methods, 326*(February), 108378. https://doi.org/10.1016/j.jneumeth.2019.108378

Hortensius, R., & Cross, E. (2018a). From automata to animate beings: The scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences, 1426*(May), 93–110.

Hortensius, R., Hekele, F., & Cross, E. S. (2018b). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems, 10*(4), 852-864. https://doi.org/10.17605/OSF.IO/UFZ5W

Hortensius, R., Kent, M., Darda, K., Jastrzab, L., Koldewyn, K., Ramsey, R. & Cross, E. S. Exploring the relationship between anthropomorphism and Theory-of-Mind in brain and behaviour. Manuscript in preparation.

Hove, M. J., & Risen, J. L. (2009). It's All in the Timing: Interpersonal Synchrony Increases Affiliation. *Social Cognition*, *27*(6), 949–960. https://doi.org/10.1521/soco.2009.27.6.949

Hsieh, T. Y., Chaudhury, B., & Cross, E. S. (2020, March). Human-Robot Cooperation in Prisoner Dilemma Games: People Behave More Reciprocally than Prosocially Toward Robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 257-259).

Hu, X., Wagley, N., & Rioboo, T. (2020). *Photogrammetry-based stereoscopic optode registration method for functional near-infrared spectroscopy. 25*(September), 1–13. https://doi.org/10.1117/1.JBO.25.9.095001

Huppert, T. J., Diamond, S. G., Franceschini, M. A., & Boas, D. A. (2009). HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Applied Optics*, *48*(10), D280. https://doi.org/10.1364/ao.48.00d280

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*. 359. 725-726.

Hyde, D. C., Aparicio Betancourt, M., & Simon, C. E. (2015). Human temporal-parietal junction spontaneously tracks others' beliefs: A functional near-infrared spectroscopy study. *Human Brain Mapping*, *36*(12), 4831–4846. https://doi.org/10.1002/hbm.22953

Irfan, B., Kennedy, J., Lemaignan, S., Papadopoulos, F., Senft, E., & Belpaeme, T. (2018). Social Psychology and Human-Robot Interaction: An Uneasy

Marriage. *ACM/IEEE International Conference on Human-Robot Interaction*, 13–20. https://doi.org/10.1145/3173386.3173389

Isaac, L., Vrijsen, J. N., Eling, P., Van Oostrom, I., Speckens, A., & Becker, E. S. (2012). Verbal and facial-emotional stroop tasks reveal specific attentional interferences in sad mood. *Brain and Behavior*, *2*(1), 74–83. https://doi.org/10.1002/brb3.38

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. https://doi.org/10.1073/pnas.1714471114

Jacobs, A. (2018). What Could Go Wrong Lessons learned when doing HRI user studies with off the shelf social robots. *Reading Teacher*, *43*(9), 666–673.

Jacoby, N., Bruneau, E., Koster-Hale, J., & Saxe, R. (2016). Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage*, *126*, 39–48. https://doi.org/10.1016/j.neuroimage.2015.11.025

Jaffe-Dax, S., Bermano, A., & Emberson, L. (2019). *Video-based motion-resilient reconstruction of 3D position for fNIRS/EEG head mounted probes*.

Jasper, H. H. (1958). The Ten-Twenty Electrode System of the International Federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 371–375.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 14.

Kahn, P. H., Ishiguro, H., Friedman, B., & Kanda, T. (2006). What is a human? - Toward psychological benchmarks in the field of human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, *3*, 364–371. https://doi.org/10.1109/ROMAN.2006.314461

Kalegina, A., Schroeder, G., Allchin, A., Berlin, K., & Cakmak, M. (2018). Characterizing the Design Space of Rendered Robot Faces. *ACM/IEEE International Conference on Human-Robot Interaction*, 96–104. https://doi.org/10.1145/3171221.3171286

Kasparov, G. (2017). *Deep thinking: Where machine intelligence ends and human creativity begins*. Hachette UK.

Kawaguchi, Y., Wada, K., Okamoto, M., Tsujii, T., Shibata, T., & Sakatani, K. (2012). Investigation of brain activity after interaction with seal robot measured by fNIRS. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 571–576. https://doi.org/10.1109/ROMAN.2012.6343812

Kennedy, J., Baxter, P., & Belpaeme, T. (2013). Social Robots for Education. *Presented at the International Summer School on Social Human-Robot Interaction, Cambridge, U.K.*, *5954*, 248116–248116.

Kewenig, V. (2018). Commentary: Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. *Frontiers in Psychology*, *9*, 2.

Kidd, C. D., Taggart, W., & Turkle, S. (2006). *A Sociable Robot to Encourage Social Interaction among the Elderly*. *May*, 3972–3976.

Kieval, H. (1997). Pursuing the Golem of Prague: Jewish Culture and the Invention of a Tradition. *Oxford University Press*, 24.

Kirsch, L. P., & Cross, E. S. (2018). *The influence of sensorimotor experience on the aesthetic evaluation of dance across the life span* (1st ed., Vol. 237). Elsevier B.V. https://doi.org/10.1016/bs.pbr.2018.03.012

Klapper, A., Ramsey, R., Wigboldus, D., & Cross, E. S. (2014). The control of automatic imitation based on Bottom–Up and Top–Down cues to animacy: Insights from brain and behavior. *Journal of cognitive neuroscience,* 26(11), 2503-2513.

Kokal, I., Engel, A., Kirschner, S., & Keysers, C. (2011). Synchronized drumming enhances activity in the caudate and facilitates prosocial commitment—If the rhythm comes easily. *PLoS ONE*, *6*(11), 1–12. https://doi.org/10.1371/journal.pone.0027272

Kompatsiari, K., Perez-Osorio, J., Davide, D. T., Metta, G., & Wykowska, A. (2018). *Neuroscientifically-Grounded Research for Improved Human-Robot Interaction*. 6.

Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D. S., & Nehaniv, C. L. (2009). Effects of Embodiment and Gestures on Social Interaction in Drumming Games with a Humanoid Robot. *Advanced Robotics*, *23*(14), 1951–1996. https://doi.org/10.1163/016918609X12518783330360

Kousta, S. (2020). Replications do not fail. *Nature Human Behavior*, *4*, 559. https://doi.org/10.1038/s41562-020-0903-0

Kruschke, J. K. (2018). *Rejecting or accepting parameter values in Bayesian estimation*. 20.

Kubota, A., Peterson, E. I. C., Rajendren, V., Kress-Gazit, H., & Riek, L. D. (2020). *JESSIE: Synthesizing Social Robot Behaviors for Personalized Neurorehabilitation and Beyond*. 10.

Kuipers, B., Feigenbaum, E. A., Hart, P. E., & Nilsson, N. J. (2017). Shakey: From conception to history. *AI Magazine, 38*(1), 88–103. https://doi.org/10.1609/aimag.v38i1.2716

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis, 1*(1), 76–85.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, 13.

Langton, S. R. H., Law, A. S., Burton, A. M., & Schweinberger, S. R. (2008). Attention capture by faces. *Cognition*, *107*(1), 330–342. https://doi.org/10.1016/j.cognition.2007.07.012

Larson, C. L., Baskin-Sommers, A. R., Stout, D. M., Balderston, N. L., Curtin, J. J., Schultz, D. H., Kiehl, K. A., & Newman, J. P. (2013). The interplay of attention and emotion: Top-down attention modulates amygdala activation in psychopathy. *Cognitive, Affective & Behavioral Neuroscience*, *13*(4), 757–770. https://doi.org/10.3758/s13415-013-0172-8

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people’s loneliness in human-robot interaction. *International Journal of Human Computer Studies*, *64*(10), 962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002

Lehmann, H., Saez-Pons, J., Syrdal, D. S., & Dautenhahn, K. (2015). In good company? Perception of movement synchrony of a non-anthropomorphic robot. *PLoS ONE*, *10*(5), 1–16. https://doi.org/10.1371/journal.pone.0127747

Leite, I., Martinho, C., & Paiva, A. (2013). Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics*, *5*(2), 291–308. https://doi.org/10.1007/s12369-013-0178-y

Li, James, Currano, R., Sirkin, D., Goedicke, D., Tennent, H., Levine, A., Evers, V., & Ju, W. (2020). On-Road and Online Studies to Investigate Beliefs and Behaviors of Netherlands, US and Mexico Pedestrians Encountering Hidden-Driver Vehicles. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 141–149.

Li, Jie, Ozog, P., Abernethy, J., Eustice, R. M., & Johnson-Roberson, M. (2016). Utilizing high-dimensional features for real-time robotic applications: Reducing the curse of dimensionality for recursive Bayesian estimation. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1230–1237. https://doi.org/10.1109/IROS.2016.7759205

Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. *Cortex*, *53*(1), 60–77. https://doi.org/10.1016/j.cortex.2014.01.013

Lockwood, P. L. (2020). Is There a 'Social' Brain? Implementations and Algorithms. *Trends in Cognitive Sciences*, 12.

Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science*, *21*(12), 1854–1862. https://doi.org/10.1177/0956797610388044

Lorenz, T., Weiss, A., & Hirche, S. (2016). Synchrony and Reciprocity: Key Mechanisms for Social Companion Robots in Therapy and Care. *International Journal of Social Robotics*, *8*(1), 125–143. https://doi.org/10.1007/s12369-015-0325-8

MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention.

Trends in Cognitive Sciences, *4*(10), 383–391. https://doi.org/10.1016/S1364-6613(00)01530-8

Martinez-Conde, S., Conley, D., Hine, H., Kropf, J., Tush, P., Ayala, A., & Macknik, S. L. (2015). Marvels of illusion: Illusion and perception in the art of Salvador Dali. *Frontiers in Human Neuroscience*, *9*(SEPTEMBER), 1–12. https://doi.org/10.3389/fnhum.2015.00496

Mehta, R. K. (2019). Neural Efficiency of Human–Robotic Feedback Modalities Under Stress Differs With Gender. *Frontiers in Human Neuroscience*, *13*, 12.

Mejia, C., & Kajikawa, Y. (2017). Bibliometric Analysis of Social Robotics Research: Identifying Research Trends and Knowledgebase. *Applied Sciences*, *7*(12). https://doi.org/10.3390/app7121316

Meltzoff, A. N. (2007). ‘Like me’: A foundation for social cognition. *Developmental Science*, *10*(1), 126–134. https://doi.org/10.1111/j.1467-7687.2007.00574.x

Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. N. (2010). “Social” robots are psychological agents for infants: A test of gaze following. *Neural Networks*, *23*(8–9), 966–972. https://doi.org/10.1016/j.neunet.2010.09.005

Miyazaki, Y., Wake, H., Ichihara, S., & Wake, T. (2012). Attentional Bias to Direct Gaze in a Dot-Probe Paradigm. *Perceptual and Motor Skills*, *114*(3), 1007–1022. https://doi.org/10.2466/21.07.24.pms.114.3.1007-1022

Mogan, R., Fischer, R., & Bulbulia, J. A. (2017). To be in synchrony or not? A meta-analysis of synchrony’s effects on behavior, perception, cognition and affect. *Journal of Experimental Social Psychology*, *72*(March), 13–20. https://doi.org/10.1016/j.jesp.2017.03.009

Molavi, B., & Dumont, G. A. (2012). Wavelet-based motion artifact removal for functional near-infrared spectroscopy. *Physiological Measurement*, *33*(2), 259–270. https://doi.org/10.1088/0967-3334/33/2/259

Morey, R., Rouder, J., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). *Package 'BayesFactor'* [R]. https://richarddmorey.github.io/BayesFactor/

Mori, M., MacDorman, K., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Mörtl, A., Lorenz, T., & Hirche, S. (2014). Rhythm patterns interaction—Synchronization behavior for human-robot joint action. *PLoS ONE*, *9*(4). https://doi.org/10.1371/journal.pone.0095195

Mubin, O. (2018). *Social Robots in Public Spaces: A Meta-review*. 8.

Müller, B. C. N., Brass, M., Kühn, S., Tsai, C. C., Nieuwboer, W., Dijksterhuis, A., & van Baaren, R. B. (2011). When Pinocchio acts like a human, a wooden hand becomes embodied. Action co-representation for non-biological agents. *Neuropsychologia*, *49*(5), 1373–1377. https://doi.org/10.1016/j.neuropsychologia.2011.01.022

Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Du Sert, N.P., Simonsohn, U., Wagenmakers, E.J., Ware, J.J. and Ioannidis, J.P., 2017. A manifesto for reproducible science. *Nature human behaviour*, *1*(1), pp.1-9.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nash, K., Lea, J. M., Davies, T., Yogeeswaran, K., & Nash, K. (2017). The Bionic Blues: Robot Rejection Lowers Self-Esteem. *Computers in Human Behavior*. https://doi.org/10.1016/J.CHB.2017.09.018

Natale, L., Bartolozzi, C., Pucci, D., Wykowska, A., & Metta, G. (2017). iCub: The not-yet-finished story of building a robot child. *Science Robotics, 2*(13), 2–4. https://doi.org/10.1126/scirobotics.aaq1026

Natale, S., & Ballatore, A. (2020). Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence: The International Journal of Research into New Media Technologies*, *26*(1), 3–18. https://doi.org/10.1177/1354856517715164

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychon Bull Rev*, 7.

Nilsson, N. J. (1969). *A Mobile Automaton: An Application of Artificial Intelligence Techniques:* Defense Technical Information Center. https://doi.org/10.21236/ADA459660

Noah, J. A., Ono, Y., Nomoto, Y., Shimada, S., Tachibana, A., Zhang, X., Bronner, S., & Hirsch, J. (2015). FMRI Validation of fNIRS Measurements During a Naturalistic Task. *Journal of Visualized Experiments*, *100*, 5–9. https://doi.org/10.3791/52116

Nomura, T., Kanda, T., & Suzuki, T. (2005). Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. *AI and Society*, *20*(2), 138–150. https://doi.org/10.1007/s00146-005-0012-7

Nordt, M., Hoehl, S., & Weigelt, S. (2016). The use of repetition suppression paradigms in developmental cognitive neuroscience. *Cortex*, *80*, 61–75. https://doi.org/10.1016/j.cortex.2016.04.002

Oberman, L. M., McCleery, J. P., Ramachandran, V. S., & Pineda, J. A. (2007). EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing, 70*(13), 2194–2203. https://doi.org/10.1016/j.neucom.2006.02.024

Oliver, D., Tachtsidis, I., & Hamilton, A. F. de C. (2018). The role of parietal cortex in overimitation: A study with fNIRS. *Social Neuroscience, 13*(2), 214–225. https://doi.org/10.1080/17470919.2017.1285812

Olmen, K. V. (2018). *Investigating the neural basis of social interaction with fNIRS*. 2017–2018.

Omer, Y., Sapir, R., Hatuka, Y., & Yovel, G. (2019). What Is a Face? Critical Features for Face Detection. *Perception, 48*(5), 437–446. https://doi.org/10.1177/0301006619838734

Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Van Overwalle, F. (2016). Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, *00*(00), 1–12. https://doi.org/10.1080/17470919.2016.1207702

Palmer, C. J., & Clifford, C. W. G. (2020). Face Pareidolia Recruits Mechanisms for Detecting Human Social Attention. *Psychological Science*, 095679762092481. https://doi.org/10.1177/0956797620924814

Pandey, A. K., & Gelin, R. (2018). A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind. *IEEE Robotics & Automation Magazine*, *25*(3), 40–48. https://doi.org/10.1109/MRA.2018.2833157

Pavlova, M. A., Galli, J., Pagani, F., Micheletti, S., Guerreschi, M., Sokolov, A. N., Fallgatter, A. J., & Fazzi, E. M. (2018). Social cognition in down syndrome: Face tuning in face-like non-face images. *Frontiers in Psychology*, *9*(DEC), 1–9. https://doi.org/10.3389/fpsyg.2018.02583

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8-13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., & Pioggia, G. (2016). Autism and social robotics: A systematic review. *Autism Research*, *9*(2), 165–183. https://doi.org/10.1002/aur.1527

Pereira, A., Oertel, C., Fermoselle, L., Mendelson, J., & Gustafson, J. (2020, March). Effects of different interaction contexts when evaluating gaze models in HRI. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 131-139).

Pereira, E. J., Birmingham, E., & Ristic, J. (2019). Contextually-Based Social Attention Diverges across Covert and Overt Measures. *Vision*, *3*(2), 29. https://doi.org/10.3390/vision3020029

Pereira, E. J., Birmingham, E., & Ristic, J. (2020). The eyes do not have it after all? Attention is not automatically biased towards faces and eyes. *Psychological Research*, *84*(5), 1407–1423. https://doi.org/10.1007/s00426-018-1130-4

Perez-Osorio, J., De Tommaso, D., Baykara, E., & Wykowska, A. (2018). Joint Action with Icub: A Successful Adaptation of a Paradigm of Cognitive Neuroscience in HRI. *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 152–157. https://doi.org/10.1109/ROMAN.2018.8525536

Perez-Osorio, Jairo, Marchesi, S., Ince, M., & Wykowska, A. (2019). More Than You Expect: Priors Influence on the Adoption of Intentional Stance Toward Humanoid Robots. *International Conference on Social Robotics*, 119–129. https://doi.org/10.1007/978-3-030-35888-4_12

Perry, J. (2019). *EnPruneChannels.* https://www.youtube.com/watch?v=aaUIl_k0yyY&t=34s

Pessoa, L., McKenna, M., Gutierrez, E., & Ungerleider, L. G. (2002). Neural processing of emotional faces requires attention. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11458–11463. https://doi.org/10.1073/pnas.172403899

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). *What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database*. *18*, 105–113.

Pinti, P., Merla, A., Aichelburg, C., Lind, F., Power, S., Swingler, E., ... & Tachtsidis, I. (2017). A novel GLM-based method for the Automatic IDentification of functional Events (AIDE) in fNIRS data recorded in naturalistic environments. *Neuroimage*, *155*, 291-304.

Pinti, P., Aichelburg, C., Gilbert, S., Hamilton, A., Hirsch, J., Burgess, P., & Tachtsidis, I. (2018). A Review on the Use of Wearable Functional Near-Infrared Spectroscopy in Naturalistic Environments. *Japanese Psychological Research*, *60*(4), 347–373. https://doi.org/10.1111/jpr.12206

Pinti, P., Devoto, A., Greenhalgh, I., Tachtsidis, I., Burgess, P. W., & de C Hamilton, A. F. (2020a). The role of anterior prefrontal cortex (area 10) in face-to-face deception measured with fNIRS. *Social Cognitive and Affective Neuroscience*, *March*, 1–14. https://doi.org/10.1093/scan/nsaa086

Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., & Tachtsidis, I. (2019). Current status and issues regarding pre-processing of fNIRS neuroimaging data: an investigation of diverse signal filtering methods within a general linear model framework. *Frontiers in human neuroscience*, *12*, 505.

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020b). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1–25. https://doi.org/10.1111/nyas.13948

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage*, *56*(4), 2356-2363.

Plichta, M. M., Herrmann, M. J., Baehne, C. G., Ehlis, A. C., Richter, M. M., Pauli, P., & Fallgatter, A. J. (2006). Event-related functional near-infrared spectroscopy (fNIRS): Are the measurements reliable? *NeuroImage*, *31*(1), 116–124. https://doi.org/10.1016/j.neuroimage.2005.12.008

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2019). *Scanning the horizon: Towards transparent and reproducible neuroimaging research*. 29.

Poldrack, R., Mumford, J., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Powell, L. J., Deen, B., & Saxe, R. (2018). Using individual functional channels of interest to study cortical development with fNIRS. *Developmental Science*, *21*(4), e12595. https://doi.org/10.1111/desc.12595

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, *84*, 320-341.

Pownall, M., Talbot, C., Henschel, A., Lautarescu, A., Lloyd, K., Hartmann, H., Darda, K., Tang, K., Carmichael-Murphy, P., & Siegel, J. (2020).

Navigating Open Science as Early Career Feminist Researchers. *Psychology of Women Quarterly*. https://psyarxiv.com/f9m47

Prescott, T. J. (2017). Robots are not just tools. *Connection Science*, *29*(2), 142–149. https://doi.org/10.1080/09540091.2017.1279125

Prescott, T. J., Epton, T., Evers, V., Mckee, K., Webb, T., Benyon, D., Conran, S., Strand, R., Buning, M. D. C., Verschure, P. F. M. J., & Dario, P. (2012). Robot Companions For Citizens: Roadmapping The Potential For Future Robots In Empowering Older People. *Proceedings of the Conference on Bridging Research in Ageing and ICT Development (BRAID).*, *May*.

Press, C. (2011). Action observation and robotic agents: learning and anthropomorphism. *Neuroscience & Biobehavioral Reviews*, *35*(6), 1410-1418.

Quadflieg, S., & Koldewyn, K. (2017). The neuroscience of people watching: how the human brain makes sense of other people's encounters. *Annals of the New York Academy of Sciences*, *1396*(1), 166-182.

Quaresima, V., & Ferrari, M. (2019). Functional Near-Infrared Spectroscopy (fNIRS) for Assessing Cerebral Cortex Function During Human Behavior in Natural/Social Situations: A Concise Review. *Organizational Research Methods*, *22*(1), 46–68. https://doi.org/10.1177/1094428116658959

Rafaeli, A., Ashtar, S., & Altman, D. (2019). Digital traces: New data, resources, and tools for psychological-science research. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721419861410

Ramsey, R. (2020). A call for greater modesty in psychology and cognitive neuroscience. *PsyArXiv*, 1–23. https://doi.org/10.31234/osf.io/hf5sv

Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prévot, L., & Chaminade, T. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B, 374*(1771), 20180033.

Reeves, B., & Nass, C. (1996). Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. *Collection Management*, *24*(3–4), 310–311. https://doi.org/10.1300/j105v24n03_14

Rennung, M., & Göritz, A. S. (2016). Prosocial consequences of interpersonal synchrony: A Meta-Analysis. *Zeitschrift Fur Psychologie / Journal of Psychology*, *224*(3), 168–189. https://doi.org/10.1027/2151-2604/a000252

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature communications*, *9*(1), 1-12.

Richardson, H., & Saxe, R. (2020). Development of predictive responses in theory of mind brain regions. *Developmental Science*, *23*(1), 1–7. https://doi.org/10.1111/desc.12863

Riddoch, K., & Cross, E. S. (2020). *"Hit the Robot on the Head… with this Mallet" – Making a Case for Including More Open Questions in HRI Research* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/4g7xc

Riek, L. D. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

Riek, L. D. (2014). The Social Co-Robotics Problem Space: Six Key Challenges. *Robotics Challenges and Vision (RCV2013).* http://arxiv.org/abs/1402.3213

Rivoire, C., & Lim, A. (2016). Habit detection within a long-term interaction with a social robot: An exploratory study. *Proceedings of the International Workshop on Social Learning and Multimodal Interaction for*

Designing Artificial Agents - DAA '16, 1-6. https://doi.org/10.1145/3005338.3005342

Ro, T., Russell, C., & Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. *Psychological Science*, *12*(1), 94–99. https://doi.org/10.1111/1467-9280.00317

Robertson, D. J., Jenkins, R., & Burton, A. M. (2017). Face detection dissociates from face identification. *Visual Cognition*, *25*(7–8), 740–748. https://doi.org/10.1080/13506285.2017.1327465

Robins, B., Dautenhahn, K., Boekhorst, R., & Billard, A. (2005). Robotic Assistants in Therapy and Education of Children with Autism: Can a Small Humanoid Robot Help Encourage Social Interaction Skills? *Universal Access in the Information Society*, *4*(2), 105–120.

Robinson, H., Macdonald, B., Kerse, N., & Broadbent, E. (2013). The Psychosocial Effects of a Companion Robot: A Randomized Controlled Trial. *Journal of the American Medical Directors Association*, 1–7. https://doi.org/10.1016/j.jamda.2013.02.007

Rorden, C. (2007). *MRIcro*.

Rosenthal-von der Pütten, A. M., Krämer, N. C., Becker-Asano, C., Ogawa, K., Nishio, S., & Ishiguro, H. (2014). The Uncanny in the Wild. Analysis of Unscripted Human-Android Interaction in the Field. *International Journal of Social Robotics*, *6*(1), 67–83. https://doi.org/10.1007/s12369-013-0198-7

Rosenthal-von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *Journal of Neuroscience*, *39*(33), 6555-6570.

Rosenthal-Von Der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., Brand, M., & Krämer, N. C. (2014). Investigations on empathy towards humans and robots using fMRI. *Computers in Human Behavior*, *33*, 201–212. https://doi.org/10.1016/j.chb.2014.01.004

Rusz, D., Bijleveld, E., & Kompier, M. (2019). Do Reward-Related Distractors Impair Cognitive Performance? Perhaps Not. *Collabra: Psychology*, *5*(1), 10. https://doi.org/10.1525/collabra.169

Rusz, D., Le Pelley, M., Kompier, M. A. J., Mait, L., & Bijleveld, E. (2020). Reward-driven distraction: A meta-analysis. *Journal of Chemical Information and Modeling*. https://doi.org/10.1017/CBO9781107415324.004

Šabanović, S. (2010). Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. *International Journal of Social Robotics*, *2*(4), 439–450. https://doi.org/10.1007/s12369-010-0066-7

Sandini, G., Metta, G., & Vernon, D. (2004). An Open Framework for Research in Embodied Cognition ∗. *International Journal of Humanoid Robotics*.

Sandini, G., Mohan, V., Sciutti, A., & Morasso, P. (2018). Social Cognition for Human-Robot Symbiosis—Challenges and Building Blocks. *Frontiers in Neurorobotics*, *12*(July), 1–19. https://doi.org/10.3389/fnbot.2018.00034

Sandygulova, A., Johal, W., Zhexenova, Z., Tleubayev, B., Zhanatkyzy, A., Turarova, A., Telisheva, Z., CohenMiller, A., Asselborn, T., & Dillenbourg, P. (2020). *CoWriting Kazakh: Learning a New Script with a Robot*. 8.

Sarrica, M., Brondi, S., & Fortunati, L. (2019). How many facets does a “social robot” have? A review of scientific and popular definitions online.

Information Technology & People, *33*(1), 1–21. https://doi.org/10.1108/ITP-04-2018-0203

Sato, S., & Kawahara, J. I. (2015). Attentional capture by completely task-irrelevant faces. *Psychological Research*, *79*(4), 523–533. https://doi.org/10.1007/s00426-014-0599-8

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., ... & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, *64*, 240-256.

Schad, D., Betancourt, M., & Vasishth, S. (2020). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods.* https://doi.org/10.1037/met0000275

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*(APR), 1–13. https://doi.org/10.3389/fpsyg.2019.00813

Scheel, A. M., Schijen, M., & Lakens, D. (2020). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *PsyArXiv*.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience 1. *Behavioral and brain sciences*, *36*(4), 393-414.

Schilbach, L., Eickhoff, S. B., Cieslik, E., Shah, N. J., Fink, G. R., & Vogeley, K. (2011). Eyes on me: An fMRI study of the effects of social gaze on action control. *Social Cognitive and Affective Neuroscience*, *6*(4), 393–403. https://doi.org/10.1093/scan/nsq067

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie canadienne*.

Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Mata Pavia, J., Wolf, U., & Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *NeuroImage*, *85*, 6–27. https://doi.org/10.1016/j.neuroimage.2013.05.004

Schreier, J. (2012). *Robot & Frank*. Sony Pictures.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. (2017). Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: A review using probabilistic atlases from different imaging modalities: Brain Anatomy Underlying Temporo-Parietal Junction. *Human Brain Mapping*, *38*(9), 4788–4805. https://doi.org/10.1002/hbm.23675

Schwartz, O. (2019, March 13). Untold History of AI: When Charles Babbage Played Chess With the Original Mechanical Turk. *IEEE SPECTRUM*. https://spectrum.ieee.org/tech-talk/tech-history/dawn-of-electronics/untold-history-of-ai-charles-babbage-and-the-turk

Selma, Š., Bennett, C. C., Chang, W., & Huber, L. (2013). *PARO Robot Affects Diverse Interaction Modalities in Group Sensory Therapy for Older Adults with Dementia*.

Senju, A., & Johnson, M. H. (2009). The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences*, *13*(3), 127–134. https://doi.org/10.1016/j.tics.2008.11.009

Shen, Q., Dautenhahn, K., Saunders, J., & Kose, H. (2015). Can Real-time, Adaptive Human-Robot Motor Coordination Improve Humans' Overall Perception of a Robot? *IEEE Transactions on Autonomous Mental Development*, *7*(1), 52–64. https://doi.org/10.1109/TAMD.2015.2398451

Shibata, T., Wada, K., & Tanie, K. (2003). Statistical Analysis and Comparison of Questionnaire Results of Subjective Evaluations of Seal Robot in Japan and UK. *Proceedings of the 2003 IEEE*, 52–57.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 Word Solution. *SSRN Electronic Journal*, 1–4. https://doi.org/10.2139/ssrn.2160588

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An Introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Simut, R. E., Vanderfaeillie, J., Peca, A., Van de Perre, G., & Vanderborght, B. (2016). Children with Autism Spectrum Disorders Make a Fruit Salad with Probo, the Social Robot: An Interaction Study. *Journal of Autism and*

Developmental Disorders, *46*(1), 113–126. https://doi.org/10.1007/s10803-015-2556-9

Skågeby, J. (2018). “Well-Behaved Robots Rarely Make History”: Coactive Technologies and Partner Relations. *Design and Culture*, *10*(2), 187–207. https://doi.org/10.1080/17547075.2018.1466567

Solovey, E., Schermerhorn, P., Scheutz, M., Sassaroli, A., Fantini, S., & Jacob, R. (2012). *Brainput: Enhancing interactive systems with streaming fnirs brain input*. 10.

Stenzel, A., Chinellato, E., Bou, M. A. T., Del Pobil, Á. P., Lappe, M., & Liepelt, R. (2012). When Humanoid Robots Become Human-Like Interaction Partners: Corepresentation of Robotic Actions. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1073–1077. https://doi.org/10.1037/a0029493

Stock, R. M., & Merkle, M. (2018). *Can Humanoid Service Robots Perform Better Than Service Employees? A Comparison of Innovative Behavior Cues*. 10.

Strait, M., Lier, F., Bernotat, J., Wachsmuth, S., Eyssel, F., Goldstone, R., & Sabanovic, S. (2020). A three-site reproduction of the joint simon effect with the NAO robot. *ACM/IEEE International Conference on Human-Robot Interaction*, 103–110. https://doi.org/10.1145/3319502.3374783

Strait, M., & Scheutz, M. (2014a). Building a Literal Bridge Between Robotics and Neuroscience using Functional Near Infrared Spectroscopy (NIRS). *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*.

Strait, M., & Scheutz, M. (2014b). Measuring users’ responses to humans, robots, and human-like robots with functional near infrared spectroscopy. *The 23rd IEEE International Symposium on Robot and Human Interactive*

Communication, 1128–1133. https://doi.org/10.1109/ROMAN.2014.6926403

Stroop R. (1935). Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology, 18*(6), 643–661.

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, *5*, 15924–15924. https://doi.org/10.1038/srep15924

Tachtsidis, I., & Scholkmann, F. (2016). False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward. *Neurophotonics*, *3*(3), 031405.

Takahashi, K., & Watanabe, K. (2013). Gaze cueing by pareidolia faces. *I-Perception*, *4*(8), 490–492. https://doi.org/10.1068/i0617sas

Takahashi, K., & Watanabe, K. (2015). Seeing objects as faces enhances object detection. *I-Perception*, *6*(5), 1–14. https://doi.org/10.1177/2041669515606007

Takayama, L., Ju, W., & Nass, C. (2008, March). Beyond dirty, dangerous and dull: what everyday people think robots should do. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 25-32). IEEE.

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(46), 17954–17958. https://doi.org/10.1073/pnas.0707769104

Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R., & Hayashi, K. (2015). Pepper learns together with children: Development of an educational application. *2015 IEEE-RAS 15th International Conference on Humanoid*

Robots (Humanoids), 270–275. https://doi.org/10.1109/HUMANOIDS.2015.7363546

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657–665. https://doi.org/10.1080/13506280500410949

Thunberg, S., Thellman, S., & Ziemke, T. (2017). Don’t Judge a Book by its Cover: A Study of the Social Acceptance of NAO vs. Pepper. *Proceedings of the 5th International Conference on Human Agent Interaction*, 443–446. https://doi.org/10.1145/3125739.3132583

Tsuzuki, D., & Dan, I. (2014). Spatial registration for functional near-infrared spectroscopy: From channel position on the scalp to cortical location in individual and group analyses. *NeuroImage*, *85*, 92–103. https://doi.org/10.1016/j.neuroimage.2013.07.025

Tulli, S., Ambrossio, D. A., Najjar, A., & Lera, F. J. R. (2019, November). Great Expectations & Aborted Business Initiatives: The Paradox of Social Robot Between Research and Industry. In *BNAIC/BENELEARN*.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, *29*(6), 11.

Urgen, B. A., Pehlivan, S., & Saygin, A. P. (2019). Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. *Neuropsychologia*, *127*, 35-47.

Vallverdú, J., & Trovato, G. (2016). Emotional affordances for human-robot interaction. *Adaptive Behavior*, *24*(5), 320–334. https://doi.org/10.1177/1059712316668238

Van Honk, J., Tuiten, A., de Haan, E., vann de Hout, M., & Stam, H. (2001). Attentional biases for angry faces: Relationships to trait anger and anxiety. *Cognition and Emotion*, *15*(3), 279–297. https://doi.org/10.1080/02699930126112

Van Honk, J., Tuiten, A., Van Den Hout, M., Koppeschaar, H., Thijssen, J., De Haan, E., & Verbaten, R. (2000). Conscious and preconscious selective attention to social threat: Different neuroendocrine response patterns. *Psychoneuroendocrinology*, *25*(6), 577–591. https://doi.org/10.1016/S0306-4530(00)00011-1

Vollmer, A.-L., Read, R., Trippas, D., & Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics*, *3*(21), eaat7111. https://doi.org/10.1126/scirobotics.aat7111

Vuilleumier, P. (2002). Facial expression and selective attention. *Current Opinion in Psychiatry*, *15*(3), 291–300. https://doi.org/10.1097/00001504-200205000-00011

Vygotsky, L. (1987). Genesis of the higher mental functions. In *The history of the development of higher mental functions. Learning to think*. Plennum.

Walbrin, J., Downing, P., & Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia*, *112*(February), 31–39. https://doi.org/10.1016/j.neuropsychologia.2018.02.023

Walbrin, J., & Koldewyn, K. (2019). Dyadic interaction processing in the posterior temporal cortex. *NeuroImage*, *198*, 296–302. https://doi.org/10.1016/j.neuroimage.2019.05.027

Wan, X., Riera, J., Iwata, K., Takahashi, M., Wakabayashi, T., & Kawashima, R. (2006). The neural basis of the hemodynamic response nonlinearity in human primary visual cortex: Implications for neurovascular coupling mechanism. *Neuroimage*, *32*(2), 616-625.

Wang, Y., & Quadflieg, S. (2014). In our own image? Emotional and neural processing differences when observing human-human vs human-robot interactions. *Social Cognitive and Affective Neuroscience*, *10*(11), 1515–1524. https://doi.org/10.1093/scan/nsv043

Waytz, A., Cacioppo, J. T., Hurlemann, R., Castelli, F., Adolphs, R., & Paul, L. K. (2019). Anthropomorphizing without social cues requires the basolateral amygdala. *Journal of cognitive neuroscience*, *31*(4), 482-496.

Weiss, A., & Bartneck, C. (2015). Meta analysis of the usage of the Godspeed Questionnaire Series. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, *2015-Novem*, 381–388. https://doi.org/10.1109/ROMAN.2015.7333568

Wesencraft, K. M., & Clancy, J. A. (2019). Using Photogrammetry to Create a Realistic 3D Anatomy Learning Aid with Unity Game Engine (Vol. 1205). *Springer International Publishing*. https://doi.org/10.1007/978-3-030-31904-5_7

Wheatley, T., Kang, O., Parkinson, C., & Looser, C. E. (2012). From Mind Perception to Mental Connection Synchrony as a Mechanism for Social Understanding. *Social and Personality Psychology Compass*, *6*(8), 589–606. https://doi.org/10.1111/j.1751-9004.2012.00450.x

Wiese, E. (2018). Seeing minds in others: Mind perception modulates low-level social-cognitive performance and relates to ventromedial prefrontal structures. *Cogn Affect Behav Neurosci*, 20.

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, *8*(OCT), 1–19. https://doi.org/10.3389/fpsyg.2017.01663

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I See What You Mean: How Attentional Selection Is Shaped by Ascribing Intentions to Others. *PLoS ONE*, *7*(9), 1–7. https://doi.org/10.1371/journal.pone.0045391

Wijeakumar, S., Huppert, T. J., Magnotta, V. A., Buss, A. T., & Spencer, J. P. (2017). Validating an image-based fNIRS approach with fMRI and a working memory task. *NeuroImage*, *147*(December 2015), 204–218. https://doi.org/10.1016/j.neuroimage.2016.12.007

Willemse, C., & Wykowska, A. (2019). In natural interaction with embodied robots, we prefer it when they follow our gaze: A gaze-contingent mobile eyetracking study. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771), 20180036. https://doi.org/10.1098/rstb.2018.0036

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684. https://doi.org/10.3758/BRM.42.3.671

Williams, E. H., Cristino, F., & Cross, E. S. (2019). Human body motion captures visual attention and elicits pupillary dilation. *Cognition*, *193*(January), 104029. https://doi.org/10.1016/j.cognition.2019.104029

Williams, E. H., & Cross, E. S. (2018). Decreased reward value of biological motion among individuals with autistic traits. *Cognition*, *171*(March 2017), 1–9. https://doi.org/10.1016/j.cognition.2017.10.017

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wodehouse, A., Brisco, R., Broussard, E., & Duffy, A. (2018). Pareidolia: Characterising facial anthropomorphism and its implications for product design. *Journal of Design Research*, *16*(2), 83–98. https://doi.org/10.1504/JDR.2018.092792

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 1–2. https://doi.org/10.1098/rstb.2015.0375

Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLoS ONE*, *9*(4). https://doi.org/10.1371/journal.pone.0094339

Yang, C.-Y., Lu, M.-J., Tseng, S.-H., & Fu, L.-C. (2017). A companion robot for daily care of elders based on homeostasis. *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 1401-1406. https://doi.org/10.23919/SICE.2017.8105748

Yang, G. Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., Nelson, B. J., Scassellati, B., Taddeo, M., Taylor, R., Veloso, M., Wang, Z. L., & Wood, R. (2018). The grand challenges of science robotics. *Science Robotics*, *3*(14). https://doi.org/10.1126/scirobotics.aar7650

Yarkoni, T. (2016). The generalizability crisis. *PsyArXiv*, 1–26.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45-57. https://doi.org/10.1109/42.906424

Zhou, L. F., & Meng, M. (2019). Do you see the ‘face’? Individual differences in face pareidolia. *Journal of Pacific Rim Psychology*. https://doi.org/10.1017/prp.2019.27

Zimeo Morais, G. A., Balardin, J. B., & Sato, J. R. (2018). FNIRS Optodes’ Location Decider (fOLD): A toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-21716-z

Zlotowski, J., & Bartneck, C. (2013). The inversion effect in HRI: Are robots perceived more like humans or objects? *ACM/IEEE International Conference on Human-Robot Interaction*, 365–372. https://doi.org/10.1109/HRI.2013.6483611

Zurn, P., Bassett, D. S., & Rust, N. C. (2020). The Citation Diversity Statement: A Practice of Transparency, A Way of Life. *Trends in Cognitive Sciences*, *24*(9), 669–672. https://doi.org/10.1016/j.tics.2020.06.009