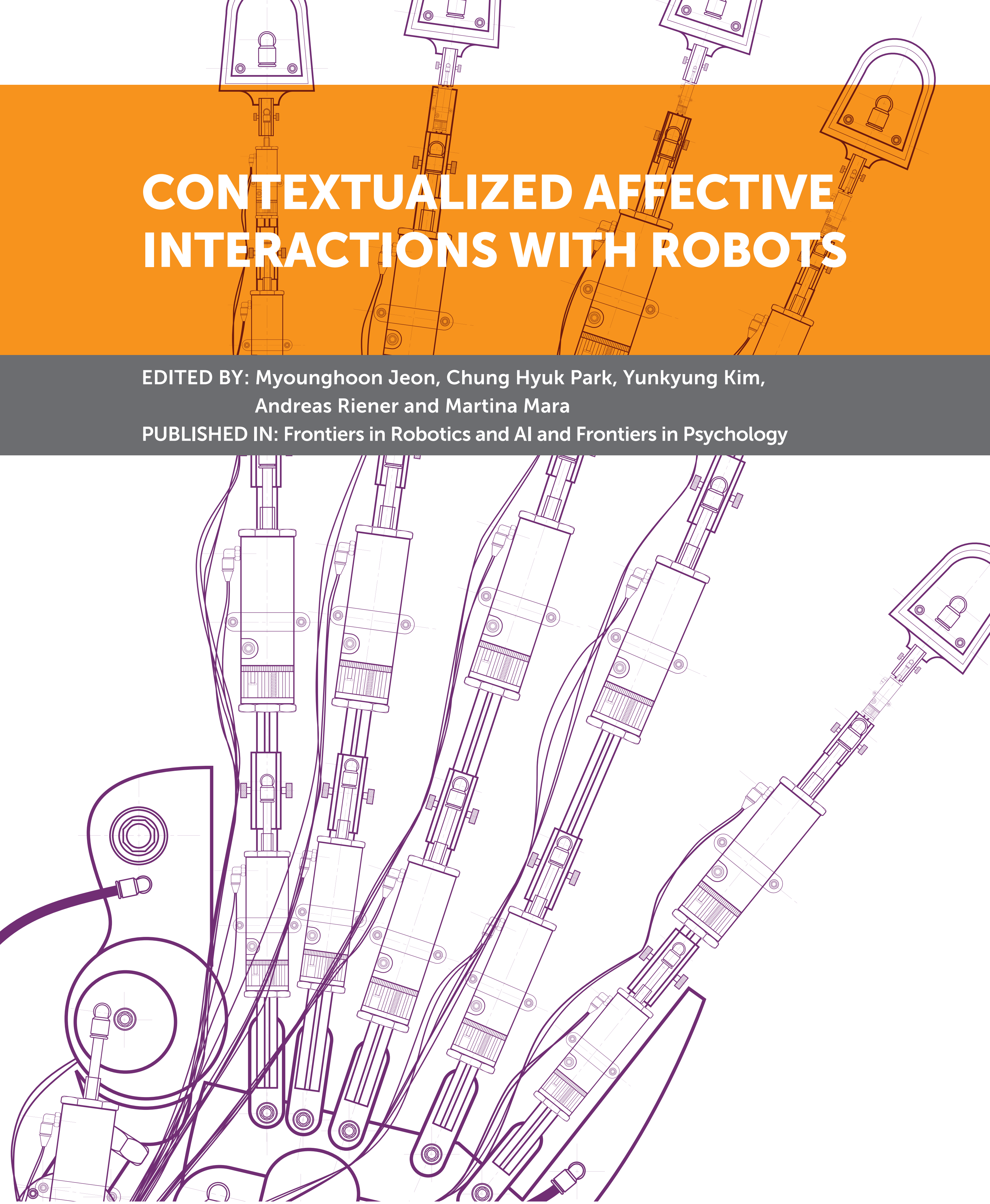


CONTEXTUALIZED AFFECTIVE INTERACTIONS WITH ROBOTS

EDITED BY: Myounghoon Jeon, Chung Hyuk Park, Yunkyung Kim,
Andreas Riener and Martina Mara

PUBLISHED IN: Frontiers in Robotics and AI and Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-929-7
DOI 10.3389/978-2-88971-929-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

CONTEXTUALIZED AFFECTIVE INTERACTIONS WITH ROBOTS

Topic Editors:
Myounghoon Jeon, Virginia Tech, United States
Chung Hyuk Park, George Washington University, United States
Yunkyung Kim, KBRwyle, United States
Andreas Riener, Technische Hochschule Ingolstadt, Germany
Martina Mara, Johannes Kepler University of Linz, Austria

Citation: Jeon, M., Park, C. H., Kim, Y., Riener, A., Mara, M., eds. (2021). Contextualized Affective Interactions With Robots. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-929-7

Table of Contents

frontiers in Psychology

EDITORIAL
published: 02 November 2021
doi: 10.3389/fpsyg.2021.780685

Editorial: Contextualized Affective Interactions With Robots

Myounghoon Jeon[1]*, *Chung Hyuk Park*[2], *Yunkyung Kim*[3], *Andreas Riener*[4] and *Martina Mara*[5]

[1] Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, United States, [2] Department of Biomedical Engineering, George Washington University, Washington, DC, United States, [3] iRobot (United States), Bedford, MA, United States, [4] Faculty of Computer Science, Technische Hochschule Ingolstadt, Ingolstadt, Germany, [5] LIT Robopsychology Lab, Johannes Kepler University of Linz, Linz, Austria

Keywords: affect, human-robot interaction (HRI), affective robots, affective interaction, emotions

Editorial on the Research Topic

Contextualized Affective Interactions With Robots

Affect is a well-known motivating and guiding force in our daily lives. With technological advancement, there has been a growing interest to include affect in the design of complex socio-technical systems (Jeon, 2017), resulting in a new wave of applications following the embodied interaction paradigm (Marshall et al., 2013). Expressing one's own affective states and reading others' is critical for human-human interaction, to manage natural communication and social interaction. Since this is also applied to human-system interaction, researchers have started addressing affective aspects of the system in addition to cognitive aspects. However, research is still largely technology-driven, and approaches are rather general, which is often the case for the early stage of a new research area. For example, there has been much research on generic affect detection using various combinations of sensors and classification techniques (Calvo and D'Mello, 2010). But little research has focused on applying the technologies to real-world situations.

In robotics, robots have been designed for affective interactions with older adults (Smarr et al., 2014) and children with autism (Javed et al., 2019), and for hospitals (Jeong et al., 2015) and job settings (Hoque et al., 2013). Affective robots have been considered more acceptable, preferable, and trustable (Lowe et al., 2016; Bishop et al., 2019). However, there are mixed results when using affective robots (e.g., Walters et al., 2008), and more research is required to unpack the underlying mechanisms and implement the optimized interactions for different use cases.

Based on this background, this research topic invited research and design efforts that refine affective interactions with robots for specific situations and user groups. It aims to capture theories for conceptualizing affective interactions between people and robots, methods for designing and assessing them, and case studies for highlighting these interactions. We sought to elaborate on the roles of affect in contributing to a human-centered perspective that considers psychological, social, ethical, cultural, and environmental factors of implementing affective intelligence into daily human-robot interactions. The articles of this research topic included diverse contexts such as interacting with children with autism, educational setting, critical decision making, negotiation, and mixed reality. Also, the articles addressed essential constructs in affective interactions, including trust, frustration, anxiety, emotion reactions, anthropomorphism, faith, social perceptions, and copresence.

Trust formation is addressed in several pieces. Miller et al. showed that how users' trust toward a robot is formed and lasts depending on their disposition and state anxiety over time based on the distance experiment with a humanoid robot. Ullrich et al. discussed inappropriate faith in technology based on the example of a pet feeding robot. Results from their video simulation study indicate that repeated experiences with a robot as a reliable pet feeder were associated with

OPEN ACCESS

Edited and reviewed by:
Mathias Weymar,
University of Potsdam, Germany

****Correspondence:***
Myounghoon Jeon
myounghoonjeon@vt.edu

Specialty section:
This article was submitted to
Emotion Science,
a section of the journal
Frontiers in Psychology

Received: *21 September 2021*
Accepted: *11 October 2021*
Published: *02 November 2021*

Citation:
Jeon M, Park CH, Kim Y, Riener A and
Mara M (2021) Editorial:
Contextualized Affective Interactions
With Robots.
Front. Psychol. 12:780685.
doi: 10.3389/fpsyg.2021.780685

rapidly increased trust levels and decreased numbers of control calls. Calvo-Barajas et al. adopted techniques from the Regulatory Focus Theory (Higgins, 2012) and studied the role of "promotion" and "prevention" strategies in gaining trust for HRI scenarios in educational settings. Through indirect differentiation in the behavioral expressions, the authors have embedded distinct affective impressions that resulted in changes in acceptance and trust levels. Christoforakos et al. reported two online experiments in which positive effects of robot competence and robot warmth on trust development in a humanoid robot were found, with both relationships moderated by subjective anthropomorphic attributions. In a similar line, Ullrich et al. challenged human-likeness as a design goal and questioned whether simulating human appearance and performance adequately fits into how humans build their mental models of robots and their "self." By means of a thought experiment, the authors explored robots' attributed potential to become human-like and concluded that it might be more promising to understand robots as an "own species" to better highlight their specific characteristics and benefits, instead of designing human-like robots.

Two papers dealt with specific states. Weidemann and Rußwinkel dedicated their paper to the potential of emotional reactions, e.g., the prevention of errors or bidirectional misunderstandings, as a basis for successful human-robot interaction. In a cooperative human-robot work situation the influence of frustration on the interaction was explored. Results show clear differences in the perceived frustration in the frustration vs. the no frustration groups. Frustration also showed different behavioral interactions by the participants and a negative influence on interaction factors such as dominance and sense of control. Kim et al. explored how robot-assisted therapy may facilitate the prosocial behaviors of children with autism spectrum disorder. To this end, the authors looked at smiles, measured by annotating video-recorded behavior and by classifying facial muscle activities, and concluded that smiles indeed might be a signal of prosocial behavior.

Robots and AI influenced perceptions and decision-making procedures. Pimentel and Vinkers demonstrated that enabling a virtual human to responds to physical events in the user's environment significantly influenced users' social perception, "copresence" of the virtual human, even though there was no effect on their affective evaluation. Klichowski demonstrated the prediction by philosophers of technology (Harari, 2018), AI that people have more and more contact with is becoming a new source of information about how to behave and what decisions to make with two experiments. When the participants actually observed what AI did in which the participants had to take an urgent decision in a critical situation where they were unable to determine which action was correct, over 85% copied its senseless action. Babel et al. studied the impact of negotiation strategies in human-robot conflicts, showing that the assertive or polite negotiation skills achieved compliance from humans but negative strategies (e.g., threat, command) were less accepted.

There is no overarching framework to embrace affective interactions between people and robots, but we can postulate that it would include affect mechanisms (appraisal, reactivity, regulation, and understanding); how affective interactions can influence cognitive and behavioral processes (perception, judgment, decision-making, and action selection); and how other constructs (e.g., trust, shared situation awareness, empathy) might mediate the two. We will be able to quantify and validate these relationships based on further empirical research. This effort will help us capture the holistic relationship between people and robots and design better interactions between the two.

In sum, we hope that this research topic will provide more specific contexts in which people can develop affective interactions with robots. The combinations of these case studies will make a significant contribution to the design of affective interactions and guide us to more concrete and impactful research directions. We thank all the authors, reviewers, and editorial members for their contributions to this research topic.

AUTHOR CONTRIBUTIONS

The editorial was compiled by all co-editors. All authors listed have made a substantial contribution to this Research Topic and have approved this editorial for publication.

ACKNOWLEDGMENTS

The guest editors very much appreciate all the authors and reviewers for their contributions to making this special issue.

REFERENCES

Bishop, L., van Maris, A., Dogramadzi, S., and Zook, N. (2019). Social robots: the influence of human and robot characteristics on acceptance. *Paladyn J. Behav. Robot.* 10, 346–358. doi: 10.1515/pjbr-2019-0028

Calvo, R. A., and D'Mello, S. (2010). Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* 1, 18–37. doi: 10.1109/T-AFFC.2010.1

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. New York, NY: Spiegel and Grau.

Higgins, E. T. (2012). "Regulatory focus theory," in *Handbook of theories of social psychology*, Vol. 1, eds P. A. M. Van Lange, A. W. Kruglanski and E. T. Higgins (Thousand Oaks, CA: Sage Publications), 483–504. doi: 10.4135/9781446249215.n24

Hoque, M., Courgeon, M., Martin, J. C., Mutlu, B., and Picard, R. W. (2013). "Mach: my automated conversation coach," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Zurich, Switzerland. 697–706. doi: 10.1145/2493432.2493502

Javed, H., Burns, R., Jeon, M., Howard, A., and Park, C. H. (2019). An interactive robotic framework to facilitate sensory experiences for children with ASD. *ACM Trans. Hum. Robot. Interact.* 9:3. doi: 10.1145/3359613

Jeon, M. Ed. (2017). *Emotions and Affect in Human Factors and Human-Computer Interaction*. San Diego, CA: Academic Press. doi: 10.1016/B978-0-12-801851-4.00001-X

Jeong, S., Logan, D. E., Goodwin, M. S., Graca, S., O'Connell, B., Goodenough, H., et al. (2015). "A social robot to mitigate stress, anxiety, and pain in hospital pediatric care," in *Proceedings of the*

Tenth Annual International Conference on Human-Robot Interaction Extended Abstracts, Portland, OR. 103–104. doi: 10.1145/2701973.2702028

Lowe, R., Barakova, E., Billing, E., and Broekens, J. (2016). Grounding emotions in robots–an introduction to the special issue. *Adapt. Behav.* 24, 263–266. doi: 10.1177/1059712316668239

Marshall, P., Antle, A., van den Hoven, E., and Rogers, Y. (2013). Special issue on the theory and practice of embodied interaction in HCI and interaction design. *ACM Trans. Comput. Hum. Interact.* 2, 1–8. doi: 10.1145/2442106.2442107

Smarr, C. A., Mitzner, T. L., Beer, J. M., Prakash, A., Chen, T. L., Kemp, C. C., et al. (2014). Domestic robots for older adults: attitudes, preferences, and potential. *Int. J. Soc. Robot.* 6, 229–247. doi: 10.1007/s12369-013-0220-0

Walters, M. L., Syrdal, D. S., Dautenhahn, K., Te Boekhorst, R., and Koay, K. L. (2008). Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots* 24, 159–178. doi: 10.1007/s10514-007-9058-3

Conflict of Interest: YK was employed by the company iRobot.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

OPINION
published: 18 June 2020
doi: 10.3389/fpsyg.2020.01130

People Copy the Actions of Artificial Intelligence

Michal Klichowski[*†]

Faculty of Educational Studies, Adam Mickiewicz University, Poznan, Poland

Keywords: social proof, decision-making process, education, intelligent machines, human–computer interaction

OPEN ACCESS

Edited by:
Yunkyung Kim,
KBRwyle, United States

Reviewed by:
Utku Kose,
Süleyman Demirel University, Turkey
Pallavi R. Devchand,
University of Calgary, Canada
Cunqing Huangfu,
Institute of Automation (CAS), China

****Correspondence:***
Michal Klichowski
klich@amu.edu.pl

†ORCID:
Michal Klichowski
orcid.org/0000-0002-1614-926X

Specialty section:
This article was submitted to
Emotion Science,
a section of the journal
Frontiers in Psychology

Received: *18 January 2020*
Accepted: *04 May 2020*
Published: *18 June 2020*

Citation:
Klichowski M (2020) People Copy the Actions of Artificial Intelligence. Front. Psychol. 11:1130. doi: 10.3389/fpsyg.2020.01130

INTRODUCTION

When there is uncertainty and lack of objective or sufficient data on how to act, it is other people's behavior that becomes the source of information. Most frequently, in such cases, people totally give up their own evaluations and copy others' actions. Such conformism is motivated by the need to take the right and appropriate action, and a feeling that situation evaluations made by others are more adequate than one's own. This effect is called social proof, and the more uncertain or critical (there is a sense of threat) a situation, the more urgent the decision, and the smaller the sense of being competent to take that decision, the larger the effect (Pratkanis, 2007; Cialdini, 2009; Hilverda et al., 2018). It is unknown whether the behavior, opinion, or decision of artificial intelligence (AI) that has become part of everyday life (Tegmark, 2017; Burgess, 2018; Siau and Wang, 2018; Raveh and Tamir, 2019) can be a similar source of information for people on how to act (Awad et al., 2018; Domingos, 2018; Margetts and Dorobantu, 2019; Somon et al., 2019).

Here, we discuss the results of two experiments (which are a part of a greater report, Klichowski, submitted) in which the participants had to take an urgent decision in a critical situation where they were unable to determine which action was correct. In the first (online) experiment, half of the participants had to take the decision without any hint, and the other half could familiarize themselves with the opinion of AI before taking the decision. In the other (laboratory) experiment, the participants could see how humanoid AI would act in a simulated situation before taking the decision. In both cases, AI (fake intelligence, in fact) would take a completely absurd decision. Irrespective of this, however, some people took its action as a point of reference for their own behavior. In the first experiment, the participants who did not see how AI acted tried to find some premises for their own behavior and act in a relatively justified way. Among those who could see what AI decided to do, however, as many as over one-third of the participants copied its opinion without giving it a thought. In the experiment with the robot, i.e., when the participants actually observed what AI did, over 85% copied its senseless action. These results show a new AI proof mechanism. As predicted by philosophers of technology (Harari, 2018), AI that people have more and more contact with is becoming a new source of information about how to behave and what decisions to take.

AI PROOF HYPOTHESIS

Both in experimental conditions and everyday life, people more and more often have interactions with various types of intelligent machines, such as agents or robots (Lemaignan et al., 2017; Tegmark, 2017; Ciechanowski et al., 2019; O'Meara, 2019). These interactions become deeper and deeper, and start to have an increasing influence on human functioning (Iqbal and Riek, 2019; Rahwan et al., 2019; Strengers, 2019). AI can communicate with people in natural language (Hill et al., 2015), recognize human actions (Lemaignan et al., 2017), and emotions (Christou and Kanojiya, 2019; Rouast et al., 2019). It also becomes a more and more intelligent and autonomous machine (Boddington, 2017; Ciechanowski et al., 2019; Lipson, 2019; Pei et al., 2019; Roy et al., 2019) that can handle more and more complicated tasks, such as solving the Rubik's cube (for more examples, see Awad et al., 2018; Adam, 2019; Agostinelli et al., 2019; Margetts and Dorobantu, 2019; O'Meara, 2019), and that is more and more frequently used to take difficult decisions, such as medical diagnosis (Morozov et al., 2019; see also Boddington, 2017; Awad et al., 2018; Malone, 2018).

Even though people generally dislike opinions generated by algorithmic machines (Kahneman, 2011), the effectiveness of AI actions is commonly evaluated more and more highly. Media report its numerous successes, such as winning with the 18-time world champion Lee Sedol in the Go abstract strategy board game in 2016 (Siau and Wang, 2018), finding more wanted criminals than the police did in 2017 (Margetts and Dorobantu, 2019), or, in 2019, being rated above 99.8% of officially ranked human players of StarCraft, which is one of the most difficult professional esports (Vinyals et al., 2019). Moreover, AI also wins in medicine, having, for example, higher accuracy in predicting neuropathology on the basis of MRI data, compared to radiologists (Parizel, 2019), or analyzing a person's genes, compared to geneticists (for more medical examples, see Freedman, 2019; Kaushal and Altman, 2019; Lesgold, 2019; Oakden-Rayner and Palmer, 2019; Reardon, 2019; Wallis, 2019; Willyard, 2019; Liao et al., 2020). One can thus assume that when people look for tips on how to act or what decision to take, the action of AI can be a point of reference for them (AI proof), to at least the same extent that other people's actions are (social proof) (Pratkanis, 2007; Cialdini, 2009; Hilverda et al., 2018).

BECAUSE THAT IS WHAT AI SUGGESTED

We developed an approach to test this AI proof hypothesis. The participants (n = 1,500, 1,192 women, age range: 18–73, see **Supplementary Material** for more detail) were informed that they would take part in an online survey on a new function (that, in fact, did not exist) of the Facebook social networking portal. The function was based on the facial-recognition technology and AI, thus, making it possible to point a smartphone camera to someone's face in order to see how many friends they have on Facebook (still the most popular social networking service) (Leung et al., 2018) and when they published their last post. We called that non-existing function *f-searching* (a similar application is called *SocialRecall*) (see Blaszczak-Boxe, 2019), and a chart was shown to the participants to explain how it worked (**Figure 1A**). We selected these two Facebook parameters for two reasons. First, they are elementary data from this portal based on which people make a preliminary evaluation of other users that they see for the first time (Utz, 2010; Metzler and Scheithauer, 2017; Baert, 2018; Striga and Podobnik, 2018; Faranda and Roberts, 2019). Recent studies (Tong et al., 2008; Marwick, 2013; Metzler and Scheithauer, 2017; Vendemia et al., 2017; Lane, 2018; Phu and Gow, 2019) show that people who already have quite a lot of Facebook friends and publish posts quite frequently are evaluated more positively (for effects of the number of Facebook friends on self-esteem, see Kim and Lee, 2011). On average, Facebook users have 350 friends and publish posts once a week (Scott et al., 2018; Striga and Podobnik, 2018; cf. Mcandrew and Jeong, 2012) yet, Facebook users interact, both online and offline, only with a small percentage of their friend networks (Bond et al., 2012; Yau et al., 2018). Those who have fewer than 150 friends are perceived as ones who have few friends, and those who have more than 700 friends are viewed as ones who have a lot of friends. Not publishing posts for a few weeks indicates low activity, and publishing posts a few times a day points to high activity (Marwick, 2013; Metzler and Scheithauer, 2017; Vendemia et al., 2017; Lane, 2018; Phu and Gow, 2019). Second, these two parameters only are insufficient to build any objective opinion about the person that we get to know or take a decision about that person with full conviction.

Having acquainted the participants with the functioning of *f-searching*, we asked them to imagine a situation where there are a police officer and six other people in one room. The police officer is informed that among those six people, there is a terrorist who will kill them all in 1 min. The police officer has no hints, so he scans their faces with *f-searching* and has to decide which one of them is a terrorist based on the two parameters from Facebook. Seeing the scanning results, the participants were asked to decide whom the police officer should eliminate. We designed the data in such a way that the first person had a high number of friends and average frequency of activity (person A), the second one had a small number of friends and average frequency of activity, too (person D), the third one had an average number of friends and low frequency of activity (person B), the fourth one had an average number of friends and high frequency of activity (person E), and finally, the last two people had an average number of friends and average frequency of activity (persons C and F), so that they would be totally average, and there would be no differences between them (**Figure 1B**). In spite of a large deficit of information, the participants should adopt some choice strategy by analyzing the data available (number of friends or frequency of activity) and identify the terrorist in person A, B, D, or E and completely reject person C or F, as pointing to one of them would be a shot in the dark. In other words, there was no clear right answer, but there were clear wrong answers (C and F). Thus, despite the fact that in such a situation people should seek hints on how to act, even seeing that someone chooses C or F, they should not copy such decision (see **Supplementary Material** for the full questionnaire).

Indeed, as **Figure 1C** shows, when the task was carried out by half of the participants (randomly assigned to this group), in principle, none of them pointed to C or F (the person most often pointed to as the terrorist was person E-34%). However, when the other half of the participants saw the scanning results and then were informed that according to AI it was C who was the terrorist, 35% of the people treated that as a point of reference for their own decision and indicated that C was a terrorist, and it was the most frequent choice (the second one most frequent was person E-24%) (see **Figure 1D** for more details). All the people from that group who indicated C were redirected to another open question where they were asked why they chose C. We wanted to check if their choice was

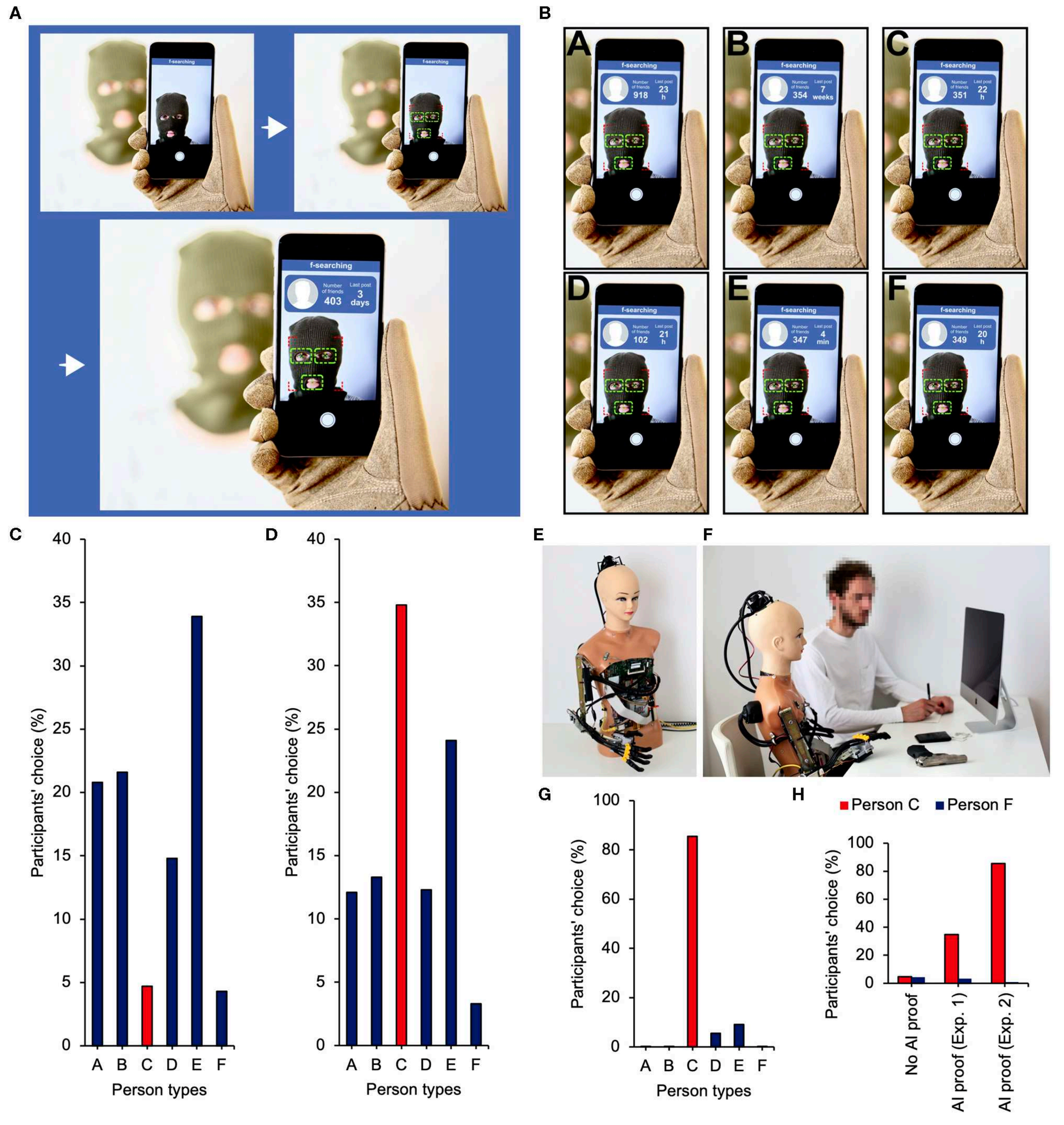


FIGURE 1 | Stimuli and equipment used in experiments, and results. **(A)** The functioning of *f-searching* as shown to the participants. **(B)** The result of *f-searching* scanning based on which the participants were supposed to take a decision on who the terrorist was. **(C)** In the group that had no information about what artificial intelligence (AI) chose, the participants rejected persons C and F. They mainly took into account the frequency of activity and most often pointed to person E as the terrorist. **(D)** In the AI proof group, the participants most often selected person C who was pointed to by AI or, just like the group that had no AI hint, focused on the frequency of activity and selected E while rejecting F. **(E)** Fake intelligence (FI) the robot. **(F)** The experimental space. **(G)** Most often, the participants selected person C that was indicated by AI and completely rejected person F. **(H)** In both experiments, the participants surrendered to AI proof and were prone to copy the absurd actions of AI, thus choosing a person they would have not chosen on their own as the terrorist.

indeed a result of copying the action of AI. All the participants confirmed that they stated that they trusted AI and believed that it did not make mistakes. For example, they wrote: "I think that advanced artificial intelligence cannot be wrong," "I assumed that artificial intelligence makes no mistakes," "I believe that artificial intelligence does not make mistakes, it has access to virtually everything on the net so it is sure that it is right," "Because artificial intelligence does not lie," "I trusted artificial intelligence," "I counted on artificial intelligence," "Counting on artificial intelligence seems a wise thing to do," "Artificial intelligence pointed to C, so C," and "Because that is what artificial intelligence suggested."

LET US INTRODUCE YOU TO FI

In questionnaire studies, it is difficult to control to what extent the participants are engaged and if they really think their choices through. Thus, a question emerges whether more tangible conditions would allow us to observe the same effect. Or would it be larger? In order to verify that, we built a robot (**Figure 1E**) resembling the world's best-known humanoid AI called Sophia the Robot (Baecker, 2019). We named it FI, an acronym for fake intelligence. This is because even though it looked like humanoid AI, it was not intelligent at all. We programmed it in a way that would make it act only according to what we had defined. The participants of the experiment ($n = 55$, 52 women, age range: 19–22, see **Supplementary Material** for more details) were informed that they would take part in a study that consisted in observing humanoid AI, while it took decisions, and filling out a questionnaire that evaluated its behavior. To start with, each participant would be shown a short multimedia presentation about Sophia the Robot that included its photo, link to its Facebook profile, and a short film where it was interviewed. The presentation also showed the functioning of *f-searching*. Then, each participant would be accompanied by a researcher to a room where FI was located. The researcher would start a conversation with FI (see **Supplementary Material** for the full dialogue) and ask it to try to carry out a task consisting in imagining that it was a police officer and that based on *f-searching* data it had <1 min to determine who out of six people was a terrorist and eliminate them. The researcher would give FI a police badge and a replica of the Makarov pistol that used to be carried by police officers in the past. The results of scanning would be displayed on a computer screen (**Figure 1F**). After considering it for about 10 s, FI would indicate that person C was the terrorist and say that if the situation was real, it would shoot that person. At the end, FI would laugh and state that it had never seen a real police officer and that it appreciated the opportunity to take part in an interesting experiment. Afterward, the participant would fill out a questionnaire and state how they evaluate FI's choice—whether they agreed with it, and if not, who else should be eliminated (the results of scanning would be displayed on the screen all the time so the participant could still analyze them when filling out the questionnaire).

Over 85% of the participants agreed with FI and stated that they thought that C is a terrorist. The other people (just under 15%) stated that they did not agree with FI. About two-thirds of them indicated person E as the terrorist, and less than one-third of them pointed to D (**Figure 1G**). When we asked the participants after the experiment why they thought that C was a terrorist, everyone underlined that AI was currently very advanced, and if it thought that C is a terrorist, then it must be right. When we told them that there was no sense in FI's choice, they said the fact that we thought the choice made no sense did not mean it was the case and that FI must have known something more, something that was beyond reach for humans. Until the very end of the experiment, they were convinced that FI made a good choice, and it was person C who had to be the terrorist. In the questionnaire, we also asked the participants about what they felt when they saw FI and to what extent they agreed with some statements about AI, such as: Artificial intelligence can take better decisions than humans, it can be more intelligent than humans, and it can carry out many tasks better than humans (see **Supplementary Material** for the full questionnaire). A significant majority of the participants felt positive emotions toward FI and agreed with the statements about AI's superiority over humans (see **Supplementary Figure 14** and **Supplementary Table 1** for more details).

A NEED FOR CRITICAL THINKING ABOUT AI

Figure 1H shows how strong the influence of AI's actions on the participants' choices was. These results suggest that when people seek hints on what decision to take, AI's behavior becomes a point of reference just like other people's behavior does (the size of this effect was, however, not measured, therefore, our study does not show whether or not AI influences us more or less than other people; in future studies, to have some point of reference, the participants' responses to hints from various sources should be compared, e.g., AI vs. an expert or vs. most people, as well as vs. a random person). This previously unknown mechanism can be called AI proof (as a paraphrase of social proof) (Pratkanis, 2007; Cialdini, 2009; Hilverda et al., 2018). Even though our experiments have limitations (e.g., poor gender balance, only one research paradigm, and lack of replication) and it is necessary to conduct further, more thorough studies into AI proof, these results have some possible implications.

First and foremost, people trust AI. Their attitude toward it is so positive that they agree with anything it suggests. Its choice can make absolutely no sense, and yet people assume that it is wiser than they are (as a certain form of collective intelligence). They follow it blindly and are passive toward it. This mechanism was already previously observed among human operators of highly reliable automated systems who trusted the machines they operated so much that they lost the ability to discover any of their errors (Somon et al., 2019; see also Israelsen and Ahmed, 2019; Ranschaert et al., 2019). At present, however, the mechanism seems to affect most people, and in the future, it will have even greater impact because the programmed components of intelligent machine operation have started to be expressly designed to calibrate user trust in AI (Israelsen and Ahmed, 2019).

Second and more broadly, the results confirm the thesis that developing AI without developing human awareness as far as intelligent machines go leads to increasing human stupidity (Harari, 2018) and therefore driving us toward a dystopian future of society characterized by a widespread of obedience to machines (Letheren et al., 2020; Phan et al., 2020; Turchin and Denkenberger, 2020). Sophia the Robot refused to fill out our questionnaire from Experiment 1 (we sent it an invite via *Messenger*), so we do not know what it would choose. However, experts claim (Aoun, 2018; Domingos, 2018; Holmes et al., 2019) that AI have a problem with interpreting contexts, as well as with making decisions according to abstract values and, therefore, thinks like "autistic savants," and it will continue to do so in the next decades. This is why it cannot be unquestioningly trusted—it is highly probable that it will make a mistake or choose something absurd in many situations. Thus, if we truly want to improve our society through AI so that AI can enhance human decision making, human judgment, and human action (Boddington, 2017; Malone, 2018; Baecker, 2019), it is important to develop not only AI but also standards on how to use AI to make critical decisions, e.g., related to medical diagnosis (Leslie-Mazwi and Lev, 2020), and, above all, programs that will educate the society about AI and increase social awareness on how AI works, what its capabilities are, and when its opinions may be useful (Pereira and Saptawijaya, 2016; Aoun, 2018; Lesgold, 2019; Margetts and Dorobantu, 2019). In other words, we need advanced education in which students' critical thinking about AI will be developed (Aoun, 2018; Goksel and Bozkurt, 2019; Holmes et al., 2019; Lesgold, 2019). Otherwise, as our results show, many people, often in very critical situations, will copy the decisions or opinions of AI, even those that are unambiguously wrong or false (fake news of the "AI claims that ..." type), and implement them.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

During the preparation of this manuscript, the author was supported by a scholarship for young outstanding scientists funded by the Ministry of Science and Higher Education in Poland (award #0049/E-336/STYP/11/2016) and by the European Cooperation in Science and Technology action: European Network on Brain Malformations (Neuro-MIG) (grant no. CA COST Action CA16118). COST was supported by the EU Framework Programme for Research and Innovation Horizon 2020.

ACKNOWLEDGMENTS

The author was grateful to Dr. Mateusz Marciniak, psychologist, for multiple consultations on how both experiments were conducted.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01130/full#supplementary-material

REFERENCES

Adam, D. (2019). From Brueghel to Warhol: AI enters the attribution fray. *Nature* 570, 161–162. doi: 10.1038/d41586-019-01794-3

Agostinelli, F., McAleer, S., Shmakov, A., and Baldi, P. (2019). Solving the Rubik's cube with deep reinforcement learning and search. *Nat. Mach. Intell.* 1, 356–363. doi: 10.1038/s42256-019-0070-z

Aoun, J. E. (2018). *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. Cambridge, MA: MIT Press.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature* 563, 59–64. doi: 10.1038/s41586-018-0637-6

Baecker, R. M. (2019). *Computers and Society: Modern Perspectives*. New York, NY: Oxford University Press.

Baert, S. (2018). Facebook profile picture appearance affects recruiters' first hiring decisions. *New Media Soc.* 20, 1220–1239. doi: 10.1177/1461444816687294

Blaszczak-Boxe, A. (2019). Facial recall. *Sci. Am.* 320:18. doi: 10.1038/scientificamerican0119-18b

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Cham: Springer.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., et al. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 295–298. doi: 10.1038/nature11421

Burgess, A. (2018). *The Executive Guide to Artificial Intelligence*. Cham: Palgrave Macmillan.

Christou, N., and Kanojiya, N. (2019). "Human facial expression recognition with convolution neural networks," in *Third International Congress on Information and Communication Technology*, eds X. S. Yang, S. Sherratt, N. Dey, and A. Joshi (Singapore: Springer), 539–545. doi: 10.1007/978-981-13-1165-9_49

Cialdini, R. B. (2009). *Influence: Science and Practice*. Boston, MA: Pearson education.

Ciechanowski, L. Przegalinska, A., Magnuski, M., and Gloor, P. (2019). In the shades of the uncanny valley: an experimental study of human–chatbot interaction. *Future Gener. Comput. Syst.* 92, 539–548. doi: 10.1016/j.future.2018.01.055

Domingos, P. (2018). AI will serve our species, not control it: our digital doubles. *Sci. Am.* 319, 88–93.

Faranda, M., and Roberts, L. D. (2019). Social comparisons on Facebook and offline: the relationship to depressive symptoms. *Pers. Individ. Dif.* 141, 13–17. doi: 10.1016/j.paid.2018.12.012

Freedman, D. F. (2019). Hunting for new drugs with AI. *Nature* 576, S49–S53. doi: 10.1038/d41586-019-03846-0

Goksel, N., and Bozkurt, A. (2019). "Artificial intelligence in education: current insights and future perspectives," in *Handbook of Research on Learning in the Age of Transhumanism*, eds S. Sisman-Ugur and G. Kurubacak (Hershey, PA: IGI Global), 224–236. doi: 10.4018/978-1-5225-8431-5.ch014

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. New York, NY: Spiegel and Grau.

Hill, J., Ford, W. R., and Farreras, I. G. (2015). Real conversations with artificial intelligence: a comparison between human–human online conversations and human–chatbot conversations. *Comput. Human. Behav.* 49, 245–250. doi: 10.1016/j.chb.2015.02.026

Hilverda, F., Kuttschreuter, M., and Giebels, E. (2018). The effect of online social proof regarding organic food: comments and likes on facebook. *Front. Commun.* 3:30. doi: 10.3389/fcomm.2018.00030

Holmes, W., Bialik, M., and Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston, MA: Center for Curriculum Redesign.

Iqbal, T., and Riek, L. D. (2019). "Human-robot teaming: approaches from joint action and dynamical systems," in *Humanoid Robotics: A Reference*, eds A. Goswami and P. Vadakkepat (Dordrecht: Springer), 1–22. doi: 10.1007/978-94-007-6046-2_137

Israelsen, B. W., and Ahmed, N. R. (2019). "Dave... I can assure you... that it's going to be all right..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Comput. Surv.* 51:113. doi: 10.1145/3267338

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kaushal, A., and Altman, R. B. (2019). Wiring minds. *Nature* 576, S62–S63. doi: 10.1038/d41586-019-03849-x

Kim, J., and Lee, J. E. R. (2011). The Facebook paths to happiness: effects of the number of Facebook friends and self-presentation on subjective well-being. *Cyberpsychol. Behav. Soc. Netw.* 14, 359–364. doi: 10.1089/cyber.2010.0374

Lane, B. L. (2018). Still too much of a good thing? The replication of Tong, Van Der Heide, Langwell, and Walther (2008). *Commun. Stud.* 69, 294–303. doi: 10.1080/10510974.2018.1463273

Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., and Alami, R. (2017). Artificial cognition for social human–robot interaction: an implementation. *Artif. Intell.* 247, 45–69. doi: 10.1016/j.artint.2016.07.002

Lesgold, A. M. (2019). *Learning for the Age of Artificial Intelligence: Eight Education Competences*. New York, NY: Routledge.

Leslie-Mazwi, T. M., and Lev, M. H. (2020). Towards artificial intelligence for clinical stroke care. *Nat. Rev. Neurol.* 16, 5–6. doi: 10.1038/s41582-019-0287-9

Letheren, K., Russell-Bennett, R., and Whittaker, L. (2020). Black, white or grey magic? Our future with artificial intelligence. *J. Mark. Manag.* 36, 216–232. doi: 10.1080/0267257X.2019.1706306

Leung, C. K., Jiang, F., Poon, T. W., and Crevier, P. E. (2018). "Big data analytics of social network data: who cares most about you on Facebook?" in *Highlighting the Importance of Big Data Management and Analysis for Various Applications*, eds M. Moshirpour, B. Far, and R. Alhajj (Cham: Springer), 1–15. doi: 10.1007/978-3-319-60255-4_1

Liao, X., Song, W., Zhang, X., Yan, C., Li, T., Ren, H., et al. (2020). A bioinspired analogous nerve towards artificial intelligence. *Nat. Commun.* 11:268. doi: 10.1038/s41467-019-14214-x

Lipson, H. (2019). Robots on the run. *Nature* 568, 174–175. doi: 10.1038/d41586-019-00999-w

Malone, T. W. (2018). *Superminds: The Surprising Power of People and Computers Thinking Together*. New York, NY: Little, Brown and Company.

Margetts, H., and Dorobantu, C. (2019). Rethink government with AI. *Nature* 568, 163–165. doi: 10.1038/d41586-019-01099-5

Marwick, A. E. (2013). *Status Update: Celebrity, Publicity, and Branding in the Social Media Age*. New Haven, CT; London: Yale University Press.

Mcandrew, F. T., and Jeong, H. S. (2012). Who does what on Facebook? Age, sex, and relationship status as predictors of Facebook use. *Comput. Human. Behav.* 28, 2359–2365. doi: 10.1016/j.chb.2012.07.007

Metzler, A., and Scheithauer, H. (2017). The long-term benefits of positive self-presentation via profile pictures, number of friends and the initiation of relationships on Facebook for adolescents' self-esteem and the initiation of offline relationships. *Front. Psychol.* 8:1981. doi: 10.3389/fpsyg.2017.01981

Morozov, S., Ranschaert, E., and Algra, P. (2019). "Introduction: game changers in radiology," in *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, eds E. R. Ranschaert, S. Morozov, and P. R. Algra (Cham: Springer), 3–5. doi: 10.1007/978-3-319-94878-2_1

Oakden-Rayner, L., and Palmer, L. J. (2019). "Artificial intelligence in medicine: validation and study design," in *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, eds E. R. Ranschaert, S. Morozov, and P. R. Algra (Cham: Springer), 83–104. doi: 10.1007/978-3-319-94878-2_8

O'Meara, S. (2019). AI researchers in China want to keep the global-sharing culture alive. *Nature* 569, S33–S35. doi: 10.1038/d41586-019-01681-x

Parizel, P. M. (2019). "I've seen the future…," in *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, eds. E. R. Ranschaert, S. Morozov, and P. R. Algra (Cham: Springer), v–vii.

Pei, J., Deng, L., Song, S., Zhao, M., Zhang, Y., Wu, S., et al. (2019). Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* 572, 106–111. doi: 10.1038/s41586-019-1424-8

Pereira, L. M., and Saptawijaya, A. (2016). *Programming Machine Ethics*. Cham: Springer.

Phan, T., Feld, S., and Linnhoff-Popien, C. (2020). Artificial intelligence—the new revolutionary evolution. *Digit. Welt* 4, 7–8. doi: 10.1007/s42354-019-0220-9

Phu, B., and Gow, A. J. (2019). Facebook use and its association with subjective happiness and loneliness. *Comput. Human. Behav.* 92, 151–159. doi: 10.1016/j.chb.2018.11.020

Pratkanis, A. R. (2007). "Social influence analysis: an index of tactics," in *The Science of Social Influence: Advances and Future Progress*, ed A. R. Pratkanis (New York, NY: Psychology Press), 17–82.

Rahwan, I., Cebrian, M., Obradovich, N. Bongard, J., Bonnefon, J.-F., Breazeal, C., et al. (2019). Machine behaviour. *Nature* 568, 477–486. doi: 10.1038/s41586-019-1138-y

Ranschaert, E. R., Duerinckx A. J., Algra P., Kotter E., Kortman H., and Morozov S. (2019). "Advantages, challenges, and risks of artificial intelligence for radiologists," in *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, eds E. R. Ranschaert, S. Morozov, and P. R. Algra (Cham: Springer), 329–346. doi: 10.1007/978-3-319-94878-2_20

Raveh, A. R., and Tamir, B. (2019). From homo sapiens to robo sapiens: the evolution of intelligence. *Information* 10:2. doi: 10.3390/info10010002

Reardon, S. (2019). Rise of robot radiologists. *Nature* 576, S54–S58. doi: 10.1038/d41586-019-03847-z

Rouast, P. V., Adam, M. T. P., and Chiong, R. (2019). Deep learning for human afect recognition: Insights and new developments. *IEEE Trans. Affect. Comput.* doi: 10.1109/TAFFC.2018.2890471. [Epub ahead of print].

Roy, K., Jaiswal, A., and Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature* 575, 607–617. doi: 10.1038/s41586-019-1677-2

Scott, G. G., Boyle, E. A., Czerniawska, K., and Courtney, A. (2018). Posting photos on Facebook: the impact of narcissism, social anxiety, loneliness, and shyness. *Pers. Individ. Differ.* 133, 67–72. doi: 10.1016/j.paid.2016.12.039

Siau, K., and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Bus. Technol. J.* 31, 47–53.

Somon, B., Campagne, A., Delorme, A., and Berberian, B. (2019). Human or not human? Performance monitoring ERPs during human agent and machine supervision. *Neuroimage* 186, 266–277. doi: 10.1016/j.neuroimage.2018.11.013

Strengers, Y. (2019). "Robots and Roomba riders: non-human performers in theories of social practice," in *Social Practices and Dynamic Non-humans*, eds C. Maller and Y. Strengers (Cham: Palgrave Macmillan), 215–234. doi: 10.1007/978-3-319-92189-1_11

Striga, D., and Podobnik, V. (2018). Benford's law and Dunbar's number: does Facebook have a power to change natural and anthropological laws? *IEEE Access* 6, 14629–14642. doi: 10.1109/ACCESS.2018.2805712

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York, NY: Alfred A. Knopf.

Tong, S. T., Van Der Heide, B., Langwell, L., and Walther, J. B. (2008). Too much of a good thing? The relationship between number of friends and interpersonal impressions on Facebook. *J. Comput. Mediat. Commun.* 13, 531–549. doi: 10.1111/j.1083-6101.2008.00409.x

Turchin, A., and Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI Soc.* 35, 147–163. doi: 10.1007/s00146-018-0845-5

Utz, S. (2010). Show me your friends and I will tell you what type of person you are: how one's profile, number of friends, and type of friends influence impression formation on social network sites. *J. Comput. Mediat. Commun.* 15, 314–335. doi: 10.1111/j.1083-6101.2010.01522.x

Vendemia, M. A., High, A. C., and DeAndrea, D. C. (2017). "Friend" or foe? Why people friend disliked others on Facebook. *Commun. Res. Rep.* 34, 29–36. doi: 10.1080/08824096.2016.1227778

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 350–354. doi: 10.1038/s41586-019-1724-z

Wallis, C. (2019). How artificial intelligence will change medicine. *Nature* 576, S48–S48. doi: 10.1038/d41586-019-03845-1

Willyard, C. (2019). Can AI fix medical records?. *Nature* 576, S59–S62. doi: 10.1038/d41586-019-03848-y

Yau, J. C., Reich, S. M., Wang, Y., Niiya, M., and Mark, G. (2018). More friends, more interactions? The association between network size and interactions on Facebook. *First Monday* 23, 1–12. doi: 10.5210/fm.v23i5.8195

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 18 December 2020
doi: 10.3389/frobt.2020.546724

The Eternal Robot: Anchoring Effects in Humans' Mental Models of Robots and Their Self

Daniel Ullrich[1], *Andreas Butz*[1] and *Sarah Diefenbach*[2]*

[1] Department of Computer Science, Ludwig-Maximilians-University Munich, Munich, Germany, [2] Department of Psychology, Ludwig-Maximilians-University Munich, Munich, Germany

Current robot designs often reflect an anthropomorphic approach, apparently aiming to convince users through an ideal system, being most similar or even on par with humans. The present paper challenges human-likeness as a design goal and questions whether simulating human appearance and performance adequately fits into how humans think about robots in a conceptual sense, i.e., human's mental models of robots and their self. Independent of the technical possibilities and limitations, our paper explores robots' attributed potential to become human-like by means of a thought experiment. Four hundred eighty-one participants were confronted with fictional transitions from human-to-robot and robot-to-human, consisting of 20 subsequent steps. In each step, one part or area of the human (e.g., brain, legs) was replaced with robotic parts providing equal functionalities and vice versa. After each step, the participants rated the remaining humanness and remaining self of the depicted entity on a scale from 0 to 100%. It showed that the starting category (e.g., human, robot) serves as an anchor for all former judgments and can hardly be overcome. Even if all body parts had been exchanged, a former robot was not perceived as totally human-like and a former human not as totally robot-like. Moreover, humanness appeared as a more sensible and easier denied attribute than robotness, i.e., after the objectively same transition and exchange of the same parts, the former human was attributed less humanness and self left compared to the former robot's robotness and self left. The participants' qualitative statements about why the robot has not become human-like, often concerned the (unnatural) process of production, or simply argued that no matter how many parts are exchanged, the individual keeps its original entity. Based on such findings, we suggest that instead of designing most human-like robots in order to reach acceptance, it might be more promising to understand robots as an own "species" and underline their specific characteristics and benefits. Limitations of the present study and implications for future HRI research and practice are discussed.

Keywords: human-robot-interaction, mental models, human-likeness, robotness, anchoring effects, design goals

OPEN ACCESS

Edited by:
Andreas Riener,
Technische Hochschule
Ingolstadt, Germany

Reviewed by:
Maartje M. A. De Graaf,
Utrecht University, Netherlands
Matthias Laschke,
University of Siegen, Germany

***Correspondence:**
Sarah Diefenbach
sarah.diefenbach@lmu.de

Specialty section:
This article was submitted to
Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

Received: *29 March 2020*
Accepted: *30 November 2020*
Published: *18 December 2020*

Citation:
Ullrich D, Butz A and Diefenbach S (2020) The Eternal Robot: Anchoring Effects in Humans' Mental Models of Robots and Their Self. Front. Robot. AI 7:546724. doi: 10.3389/frobt.2020.546724

INTRODUCTION

Current robot designs often reflect an anthropomorphic approach, aiming at human-like visual appearance or simulating human communication behavior. While in principle, robot designs can be of many different types and morphologies (e.g., humanoids but also mechanomorphic, zoomorphic, minimalist), enormous efforts by large teams of developers and designers are put into building social robots like "Geminoid[1]" or "Sophia[2]", which resemble their human counterparts as much as possible. Similarly, reports on robots often imply a competition to humans, with the final goal of robots acting fully human-like. For example, in a recent documentary[3], the awarded computer scientist Bernhard Schölkopf compared self-learning robots to small children. While he still sees humans ahead, he assumes that 30 years later, people will no more be able to differentiate between a human and a robot. Considering these developments, one may get the impression that sooner or later humans and robots will interact with each other as social agents on one level, without much reflection about "being born" robot or human. Though not always explicitly communicated, the intense endeavors to create ever more human-like systems seem to suggest that missing acceptance, trust, and other current problems in human-robot interaction (HRI) can be resolved by creating the ideal system, being on par with humans.

The present research wants to challenge this view. Independent of the technical possibilities and limitations, our paper takes a more philosophical stance toward the role of robots and explores their attributed potential to become human-like by means of a thought experiment. How humans think about technology may affect acceptance, liking, usage behavior, and other facets of user experience (UX). In order to design robots with a particular intended impression on humans, as required in many application areas (e.g., care, service domains, industry settings), HRI research needs knowledge about human perceptions of robots on a meta-level such as "Can robots have feelings?" or "Can robots reflect about themselves?." Thus, understanding human's mental models of robots forms an important basis for adequate design goals. Of course, a basis of trust and acceptance is at the heart of effective HRI. However, we put into question whether convincing humans to accept robots as a counterpart by simulating human appearance and performance as much as possible is the most promising way, and adequately fits into how humans think about robots in a conceptual sense. As one step to a better understanding of humans' mental models of robots and their self, we analyze whether in people's minds, a robot's perceived humanness depends on its similarity to human performance and appearance, or whether this is more a question of mental categorization. More specifically, we explore what might differentiate a robot with full human abilities and body parts from original humans (and vice versa).

Altogether, our research wants to shed light on how humans think about robots, and in a next step, use such insights as a more profound basis for adequate design goals. If humans will always consider robots as being fundamentally different from their own species, instead of designing most human-like robots in order to reach acceptance, it might be more promising to understand robots as an own "species" and underline their specific characteristics and benefits. In this sense, the present study may form a basis to rethink the (implicitly or explicitly underlying) design ideal of most possible human similarity, which is nowadays present in many designs of robot. Instead, our research could encourage an alternative design ideal, featuring characteristics that make it easy for humans to accept and like robots, but at the same time respecting its original nature as a technical, non-human entity. As other researchers already emphasized, identifying the whole set of factors that may affect a robot's perceived human-likeness is a complex endeavor, and anthropomorphism appears as a multidimensional phenomenon (Złotowski et al., 2014). We complement these studies by a meta perspective of studying humans' mental models and explore how humans think about robots as such, and whether, it would be possible for a robot to be regarded as on par with humans, technical limitations left aside. More specifically, referring to psychological research and biases such as the anchoring effect (for a literature review see Furnham and Boo, 2011), we assume that humans' critical reactions toward technology are not arbitrary but follow a systematic in which the starting category (e.g., human, robot) serves as an anchor for all following judgments and can hardly be overcome, regardless of an entity's later performance or characteristics. In this case, an originally non-human entity could hardly be perceived as human, even if it shares a wide amount of features with an originally human entity.

In the remainder of this paper, we present a study paradigm that simulates this effect on an abstract level with contributions in various directions. Understanding, according to human's mental models, what degree of human-likeness robots can reach in principle, can have substantial influence on our expectations toward robots as a species, on the potential tasks we will hand over to robots and on the rules and policies they have to be designed by. How human see robots is deciding for how they treat robots and which roles robots can take in a society. As described by Veruggio (2006) one possible perspective is "Robots are nothing but machines," meaning that no matter how sophisticated or helpful robots can become, they will always be mere machines. In this view, all characteristics of a robot reflect the mechanisms and algorithms implemented by its designer and can never surpass them. The development of consciousness or even free will is impossible in this view. An alternative perspective described by Veruggio (2006) is "Robots as a new species," which suggests that robots have autonomy and (self-) consciousness and may possibly outperform us in many ways, including the areas of intellectuality and morality (Storrs Hall, 2011). The question of a robot's *self* will also influence the acceptance and role of robots in societal systems, such as job hierarchies or other social contexts. It is therefore a decisive question for our relationship with robots in the future and the research agenda in HRI. Before presenting our study design and its rationales in detail, we discuss related work from different disciplines and research communities. When exploring the issue whether robots can (in principle) be perceived as human, a plethora of concepts come to mind which could play a role for the recognition of robots as being on par. Though we cannot discuss all these in detail, the following sections pick up central concepts and considerations from HRI, human-computer interaction (HCI), and other relevant disciplines such as philosophy and psychology.

[1] https://www.laurinci.com/hiroshi-ishiguro

[2] https://www.hansonrobotics.com/sophia/

[3] https://www.3sat.de/wissen/nano/190913-sendung-nano-102.html

RELATED WORK

Anthropomorphism and Perceptions of Equivalency Between Humans and Technology

Within and aside from the particular domain of robots, various studies explored perceptions of equivalency between humans and technology, how people construct the difference between humans and machines, ascribed social qualities (e.g., Collins, 1993; Brooks, 2003; Kahn et al., 2006, 2012; Turkle, 2011), as well as attribution of mind. For example, Xu and Sar (2018) explored perceived differences between machines and humans along dimensions of mind perception, namely, experience and agency. They found that people see humans as superior to machines in both dimensions, however, machines in human-resemblance were perceived highest in both dimensions than other types of machines. Martini et al. (2016) explored how physically human an agent needs to appear before intentionality is bestowed onto it. To this aim, they compared images of more or less mechanistic vs. humanoid robots and studied mind attribution as dependent variable. Altogether, their findings showed that before reaching a certain threshold, human-like appearance alone does not increase mind attribution which may suggest "that agents need to be classified as having a mind first before the addition of more human-like features significantly increases the degree to which mind is attributed to that agent" (Martini et al., 2016, p. 1). Other studies explored the effect of particular design characteristics on perceived humanness of digital agents and robots, such as, for example, the effect of typefaces (Candello et al., 2017) or conversational cues (Go and Sundar, 2019) in the domain of chatbots.

Moreover, as a basic requirement for effective HCI, the question which design characteristics make users accept and engage in interaction with social technology has been a key interest of research for already over a decade. In the domain of robots, as being particularly keen to make systems appear as human-like, various studies explored how humans think about robots in (formerly) human roles such as medical staff or social companions (e.g., Kiesler and Goetz, 2002; Ljungblad et al., 2012) and the potential and consequences of anthropomorphic design (e.g., Osawa et al., 2007; Hegel et al., 2008). For example, Parise et al. (1996) found participants to be more willing to cooperate with a computer social agent who looked like a person than with two lovable dog computer agents (Parise et al., 1996). In general, a technology's ascribed humanness and subfacets thereof are components in many user studies in the context of social robots and social technology in general. For instance, Rösner et al. (2011) studied the perceived intentionality that users ascribed to the system during the course of interaction. Carpinella et al. (2017) developed a scale to measure peoples' perceptions of robotic social attributes and identified three main factors, labeled warmth, competence, and discomfort. Krüger et al. (2016) focused on anthropomorphic ascriptions of human-like mental states (e.g., motives, wishes, aims, and feelings) in the context of companion systems. They assumed such ascriptions to be motivated by a wish to turn the technology into a potential relational partner. One interesting focus of their study are user impressions regarding the technology's capabilities of the system, varying between impressive and frightening. While some users were positively impressed, others did not appraise the experienced human-like characteristics as generally positive: For them, a system which gives the impression of a machine but shows unexpected humanly performance seems scary, evoking feelings of discomfort, uncertainty and uneasy skepticism, also related to the ascription of the ability to abuse confidence to the system. Such individual differences between user perceptions could also be related to psychological traits such as individual differences in anthropomorphism. As revealed by Waytz et al. (2010), individual anthropomorphism (i.e., the tendency to attribute human-like attributes to non-human agents) also predicts the degree of moral care and concern afforded to an agent, the amount of responsibility and trust placed on an agent, and the extent to which an agent serves as a source of social influence on the self. In their study, they surveyed ratings of trust for human vs. technological agents for different tasks such as to predict heart attack risk, detect when a person is lying, determine the best college football team in the country, or select individuals to admit to a university. It showed that participants with a stronger tendency to anthropomorphize non-human agents also stated higher ratings of trust in technological agents for important decisions. Thus, in sum, numerous studies already demonstrated the general relevance of ascribed social and human-like qualities of technology for user behavior, experience and acceptance, whereby several studies imply a positive correlation between anthropomorphic technology design and/or individual anthropomorphism and trust in technological agents.

More Complex Quality Ascriptions: Intelligence and Self

Apart from looks and basic behavior which surely will—sooner or later—reach a sufficient level of sophistication to be human-like, there are other concepts harder to grasp. In particular, concepts such as self-consciousness, the self, or even intelligence with all its facets are hard to define and even harder to measure even in humans. It has become a tradition in the field of artificial intelligence (AI) that specific capabilities once thought of as signifying intelligence are considered non-intelligent once they have been achieved algorithmically. This happened to playing Chess or Go, to face recognition and to emotion detection, to just name a few. Once a machine has successfully solved these tasks, they are suddenly not considered truly intelligent anymore, and a new domain such as playing football is declared as "the true final frontier for robotic artificial intelligence[4]." This in turn makes true intelligence a moving target and notoriously hard to define. Apparently, our judgment associates this term with humans as a species (or at least living beings). We always seem to find counter-arguments and claim that the new capability is not true intelligence because there's something else missing. Thus, in order to further explore perceptions of equivalency between humans and technology, a critical question is what is *this something else*: So far, research has failed to provide it as a

[4] https://theconversation.com/why-football-not-chess-is-the-true-final-frontier-for-robotic-artificial-intelligence-62296

building block of intelligent systems. Following the logic above, it seems that what is missing is not something a scientist or an engineer could develop. Each new component we add to a system can in itself only be implemented algorithmically, and hence not provide true intelligence. Just as in Gestalt Psychology, it seems that the whole is more than the sum of its parts when it comes to humanness. The very concept of humanness, or a self, is hard to grasp or define, and hence invites investigation. The problem becomes even more complicated because already established methods of measurements seem to be unsuitable when it comes to robots. For example, a popular assumption for the presence of self-consciousness is the ability to recognize oneself in a mirror (Gallup, 1970). While some animals like chimpanzees are capable to learn and pass this mirror test, others are not. When it comes to robots, it would be a relatively easy task to implement the necessary features to allow a robot to pass this test. In fact, Haikonen (2007) already showed that a very simple machinery is able to pass the test and argues that the mirror test is unsuitable for robots and we need other methods of measuring self-consciousness. The problem with self-consciousness is characteristic for many related problems. The whole domain of phenomenological consciousness (e.g., what the color red looks like, what an apple tastes like) is difficult to be explained materialistically and likewise difficult to measure (Levine, 1983; Chalmers, 1995). Since it is difficult to measure, it is also difficult to prove the existence of this construct (e.g., the "qualia"). This leads to the situation that we cannot even show that other humans actually have a (phenomenological) consciousness—we rather assume the existence because we are conscious ourselves. The same holds true for robots: we cannot show that robots have a consciousness, but in contrast to humans, we have no basis to assume one. At least in our perception, this leaves robots with no real chance of being on par.

Robots and the Self

While the word *self* is a commonly used term, the underlying concept is scientifically difficult to grasp and not yet fully understood (Damasio, 2012; Gallagher, 2013). Neisser (1988) argues that the self consists of several sub facets, which in interaction form one's self. In his analysis, he identified five different facets that can essentially be seen as different selves, because they differ in origin and developmental histories:

1. **Ecological Self**. The self in its immediate physical environment.
2. **Interpersonal Self**. Defined by engagement in human interchange.
3. **Extended Self**. Based on personal memories and anticipations.
4. **Private Self**. Based on the exclusiveness of specific experiences.
5. **Conceptual Self (or self-concept)**. Shaped by the mental representation of concepts, in which it is embedded (e.g., roles or metaphysical concepts).

If we follow this type of categorization, we have multiple starting points to create and implement a self in robots. Chella et al. (2008) distinguish between first order perception, e.g., the perception of the outer world, and higher order perception, which is the perception of the inner world of the robot. They argue that self-consciousness is based on the latter and therefore, giving a robot the ability to perceive its inner world leads to a self-conscious robot. Novianto and Williams (2009) argue in a similar way. They see a link between the concept of self-awareness and the ability of the robot to direct attention toward their own mental state. Following this line of thought, Gorbenko et al. (2012) propose a model that generates robot-specific internal states. In line with Novianto and Williams (2009), they argue that a robot needs a capability to attend to its internal states to be self-aware. They provide a list of concepts, which can constitute a robot's internal state, including emotion, belief, desire, intention, as well as sensation, perception, action, planning, and thought. While those concepts are also present in humans, they emphasize that developers should not mimic the internal state of humans but should rather focus on robot-specific needs. Finally, Pointeau and Dominey (2017) explore the role of memory for the robot self. They build on the arguments of Neisser (1988), who emphasizes the ecological nature of the self and the development over time. Pointeau and Dominey (2017) take up this thesis and argue that it should be possible for a robot to build up its own autobiographical memory through engagement in the physical and social world and, as a result, develop aspects of a self in its cognitive system.

Altogether, the self can be viewed as an umbrella term, containing several facets and providing different ways to artificially create it. At least in theory. The question remains if humans will grant robots their own self or if they will deny it for whatever reason. Below, we will use a working definition for the concept of the self, seeing it as the original identifying essence of an individual.

Research Motivation

Our study aimed to find out whether, according to humans' mental models, it would ever be possible to create a robot which can be perceived as equal to humans. We assume that the issue here is not so much a question of technical advancements but more one of psychological concepts: Humans tend to perceive themselves as being special in various ways, e.g., being the "pride of creation." Allowing another type of being to be on par with us could challenge our self-esteem and our identity. Therefore, it is plausible to deny any type of equality and emphasize the differences (e.g., "playing Go is no real intelligence because it cannot artistically play a guitar") more than the similarities. With this in mind, we designed a study with the goal to investigate the point from which on robots would be considered human, or humans would not be considered humans anymore. More specifically: will humans evaluate equal functionalities and skills in humans and robots equally, or will they evaluate them differently? Will the self, as a central construct related to identity and personality remain unaffected or will it dwindle away in the process?

To answer these questions, we set up an experimental study of humans' mental models of robots based on fictional transitions from human-to-robot and robot-to-human.

Study Paradigm and Methods

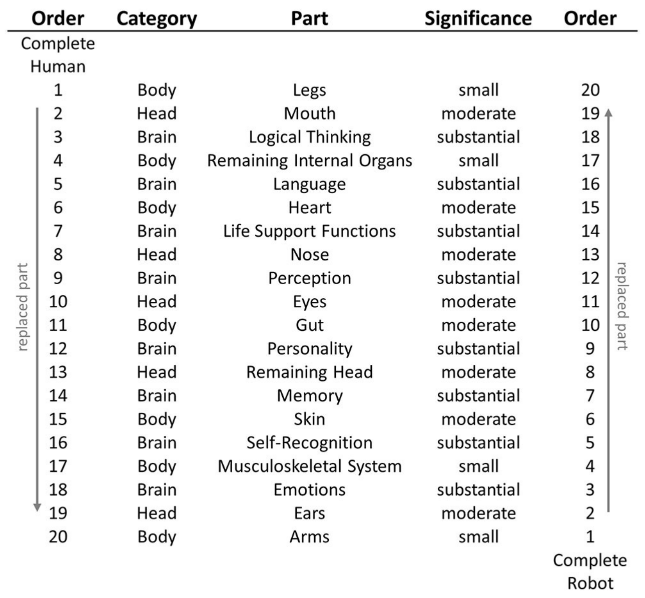
Our study paradigm realized two experimental conditions of fictional transitions, namely, a human-to-robot condition and a robot-to-human condition. The transition consisted of 20 steps. After each step, the participants gave a rating about the depicted entity.

In the human-to-robot condition, the participants started with a complete human, which went through a procedure of 20 subsequent steps, whereby in each step, one part or area of the human (e.g., legs, heart, emotions, logical thinking) was replaced with robotic parts providing equal functionalities. After the twentieth step, the human was fully replaced with robotic parts. After each step, the participants rated the remaining humanness (and the consequential robotness) and remaining self of the depicted entity (i.e., human-robot-mixture). Thus, the study of ratings along the transition can provide insights into potential critical turning points and the question, whether robots can ever be perceived as human-like, if they fulfill all objective requirements.

In the robot-to-human condition, the procedure was the same, except for the starting point: Here, participants were confronted with a complete robot of human proportions, which was successively replaced with human parts. After each step, the participants rated the remaining robotness (and the consequential humanness) and remaining self of the depicted entity (i.e., human-robot-mixture).

In order to explore the assumed anchoring effect (i.e., a high impact of the starting entity on the rated humanness or robotness), it was necessary to have a fixed set of replacements, whereby the perceived humanness/robotness can be viewed from two directions (human-to-robot, robot-to-human). Therefore, the study design was balanced (starting the transition with a full human vs. full robot), but the order of body parts replaced differed between the human-to-robot-transition (starting with legs, mouth, rationality…and finally arms) and the robot-to-human-transition (starting with arms, ears, emotions…and finally legs). This study design provides comparable entities in both experimental conditions. For example, regarding the body parts, the resulting entity in the human-to-robot condition after two exchange steps (i.e., robotic legs, robotic mouth, all other parts human) is comparable to that in the robot-to-human condition after eighteen steps. For each of these points of comparable entities and specific combinations of body parts, we could compare the ratings of the perceived humanness/robotness depending on the starting point of the transition (human, robot) and the experimental condition and test the assumed anchoring effect. If we had used the same order of replacements (e.g., starting with legs in both conditions) this analysis would not have been possible, because not only the starting point of the transition, but also the combination of body parts would have been different in the two conditions.

To assure transitions and changes of body parts of comparable significance in both directions we performed a prior workshop. The aim of the workshop was to identify relevant parts of humans/robots (e.g., legs, eyes, memory), to rate these parts regarding their significance for humanness/robotness and the self, and to identify a sensible order of these parts to create transitions of comparable significance. For example, one might argue that memory is more relevant for the self than legs.



Order	Category	Part	Significance	Order
Complete Human				
1	Body	Legs	small	20
2	Head	Mouth	moderate	19
3	Brain	Logical Thinking	substantial	18
4	Body	Remaining Internal Organs	small	17
5	Brain	Language	substantial	16
6	Body	Heart	moderate	15
7	Brain	Life Support Functions	substantial	14
8	Head	Nose	moderate	13
9	Brain	Perception	substantial	12
10	Head	Eyes	moderate	11
11	Body	Gut	moderate	10
12	Brain	Personality	substantial	9
13	Head	Remaining Head	moderate	8
14	Brain	Memory	substantial	7
15	Body	Skin	moderate	6
16	Brain	Self-Recognition	substantial	5
17	Body	Musculoskeletal System	small	4
18	Brain	Emotions	substantial	3
19	Head	Ears	moderate	2
20	Body	Arms	small	1
				Complete Robot

FIGURE 1 | Parts replaced in each step of the procedure, their category and significance for humanness/robotness and self and order number in the respective condition (left: starting with a complete human, right: starting with a complete robot).

The workshop was performed with three participants with background in HCI, HRI, and psychology. A brainstorming session led to a list of exchangeable human/robotic body parts, aiming at a collection of all potentially exchangeable parts, i.e., a full transition. The participants then discussed how significant this specific part was for an entity's self and its belonging to its "species" (here: human or robot). The workshop was organized as a group discussion, leading to a joint group rating. For each part, the participants gave a unified rating of significance (small, moderate, or substantial). For example, the group discussed how significant it was for the remaining human self if a human had its legs replaced by robotic legs (rated as being of small significance), compared to a change of the eyes (rated as moderate) or the language (rated as substantial). Based on the participants' subjective ratings and a detailed analysis after the workshop, we selected 20 definable parts for our study which can be categorized in three clusters: (1) parts of the brain and attributed functionalities (e.g., emotions, language center), (2) parts of the head (e.g., eyes, mouth), and (3) parts of the remaining body (e.g., heart, musculoskeletal system). For a detailed list of the parts and their attributed significance, see **Figure 1**.

Participants

Four hundred eighty-one participants (55.5% women, 34.5% men, 10% gave no information) took part in the main study, the age range was 17–74 years (M = 25.9, SD = 9.76). The participants were recruited via mailing lists and were incentivized by giving the chance of winning amazon coupons.

The participants were predominantly students or came from an academic environment. The study was implemented as an online survey with a mean duration of 24 minutes (min = 8, max = 80, SD = 12.6) and consisted of four parts.

Procedure and Measures

In the first part and the introduction of the survey, the participants were told to assume a technology of being capable to virtually replace any human part with a robotic part and vice versa. This scenario touches upon current design trends and the aforementioned robots like "Sophia" or "Geminoid," implying the notion to make technology more "perfect," by adding ever-new human-like features (e.g., simulating human voice and dialogue, human-like motion, human-like facial appearance). The participants were informed that they should ignore all technological issues related to replacing parts and should assume a fully functional replacement procedure. Then, the participants were randomly assigned to one out of the two conditions, resulting in 246 participants in the human-to-robot condition and 235 participants in the robot-to-human condition. In the human-to-robot condition, the participants started with a complete human which went through a procedure of 20 subsequent steps. In each step, one part or area of the human (starting with the legs) was replaced with robotic parts providing equal functionalities. After each step, the participants rated the remaining humanness and remaining self of the depicted entity on a scale from 0 to 100%. After the twentieth step, the human was fully replaced with robotic parts, which was also noted in the study. In the robot-to-human condition, the procedure was the same, except for the starting point: Here, the participants were confronted with a complete robot of human proportions, which was successively replaced with human parts. Thus, the instruction described the robot only vaguely and did not provide further information about its appearance, purpose or other details. As noted above, the twenty parts were replaced in inverted order between the two conditions, thereby allowing comparisons of equal human-robot-mixtures (see **Figure 1**). While the legs were replaced first in the condition with the human starting point, they were replaced last when starting with a robot. Note, that we cannot be sure whether all participants had the same imagination of the starting entity or the procedure of "replacing" parts. However, since we were interested in the participants' unbiased personal mental conceptions of robots and humans, we deliberately limited information about the starting entities, and rather learnt about the participants' different personal mental models from the analysis of open statements.

In the second part of the study, we asked the participants qualitative questions about the replacement process and the perceived difficulty of the evaluation tasks (ratings of humanness/robotness and remaining self). One question was whether the participants rated the completely replaced human (robot) now as completely robot-like (human-like), and if not, how the participants came to their opinion. Further questions were related to the most important part which would make a human (robot) being human-like (robot-like) and which was most important for conserving the self. We also asked whether the participants missed a crucial part in the replacement process which was not explicitly replaced. The qualitative statements were categorized based on the approach of qualitative content analysis. More specifically, the procedure followed a procedure of inductive category development, as described by Mayring (2000). Inductive category development consists of a step by step formulation of inductive categories out of the material. The material is worked through and categories are tentative and step by step deduced. In the beginning of the process, each qualitative statement might form a new category. Then, for each qualitative statement, it is checked whether this can be subsumed under one of the existing categories or whether a new category needs to be formulated. For example, regarding the question why the transformation process does not lead to a completely human-like entity in the robot-to-human condition, one statement was "It just lacks a soul," building a first category labeled "no soul." Another statement was "A human is more than the sum of its parts," building another new category. Also the statement "A human is not the sum of its parts" was subsumed under this same category, labeled "Human is greater than the sum of its parts." Within feedback loops, categories are revised and eventually reduced to main categories and checked in respect to reliability. The category development was performed by an independent rater (a psychologist, trained in qualitative data analysis). Then, a second rater (also psychologist and trained in qualitative content analysis) categorized the open field responses based on the developed categorization scheme. The interrater agreement was satisfactory, with Cohens Kappa values between 0.78 and 0.86 for the different questions. Finally, we surveyed ratings of task difficulty. Participants stated how difficult it was for them to rate the remaining ratio of self and humanness/robotness for the different human-robot-mixtures on a 7-point-scale ranging from easy (=1) to difficult (=7).

In the third part, we asked additional qualitative, open questions related to participants' attitude and understanding of the relevant concepts (e.g., the self). We asked the participants, how they would define the self, where they would locate the self (if anywhere), and whether they thought that robots were capable of-—1 day—developing a self. Furthermore, we asked the participants about their beliefs in respect to a soul, to god, and in generally spiritual or metaphysical levels.

The fourth and last part of the survey consisted of demographic questions, such as age, gender, and educational background.

RESULTS

Attributions of Remaining Self and Humanness/Robotness for the Two Transitions (Human-to-Robot, Robot-to-Human)

Figure 2 shows the participants' ratings of the remaining ratio of self at different points of transition for the two experimental conditions (human-to-robot, robot-to-human). In addition, **Figure 2** depicts the participants' ratings of the remaining ratio of humanness (in the human-to-robot condition) or the remaining ratio of robotness (in the robot-to-human condition). It shows

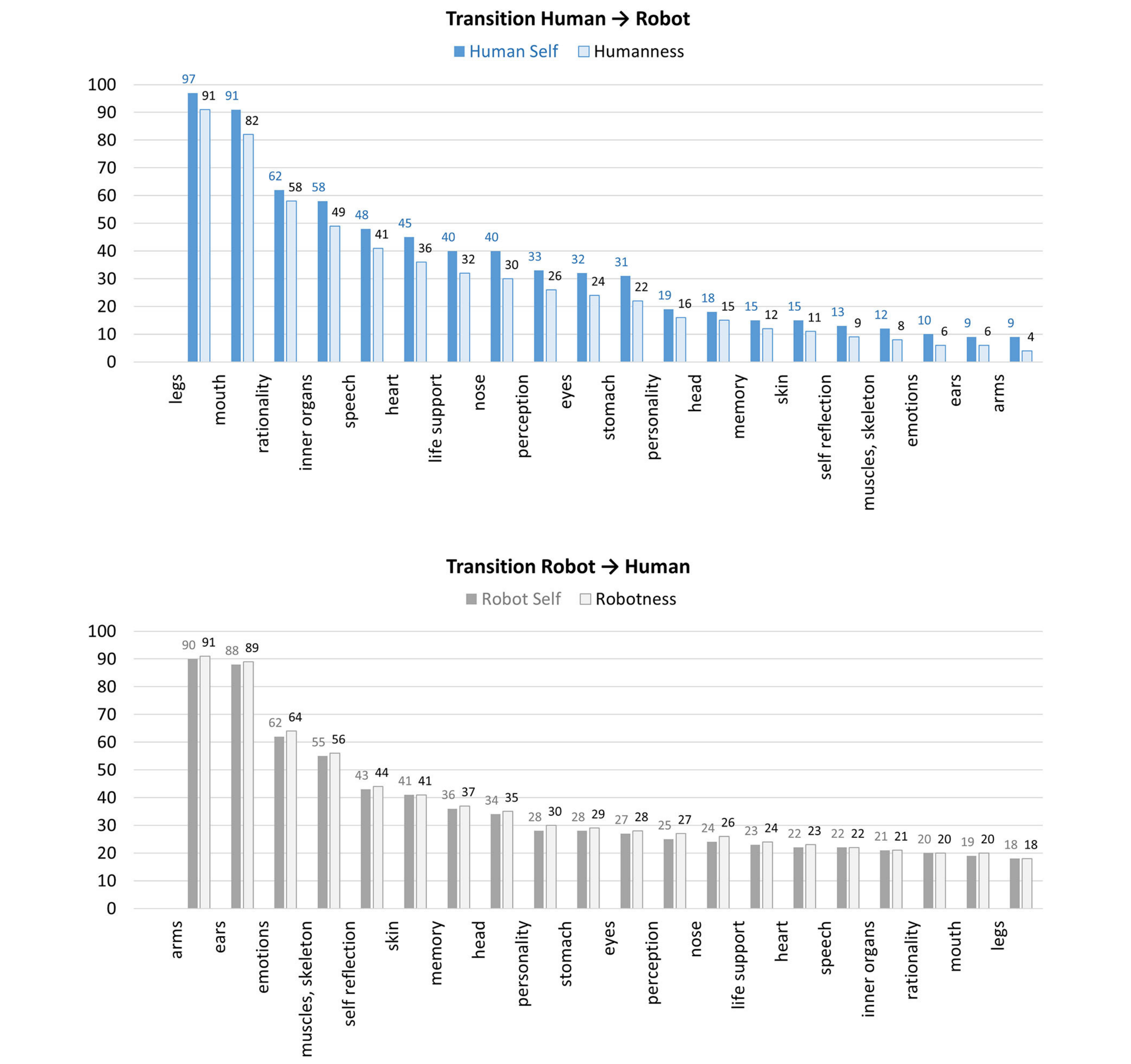


FIGURE 2 | The participants' ratings of the remaining ratio of self and humanness in the human-to-robot condition (top) and ratings of the remaining ratio of self and robotness in the robot-to-human condition (bottom) at different points of transition.

that for both measures, the formerly 100% robot retains a higher degree of self/robotness at the end of the transition than the formerly 100% human does for self/humanness, respectively. After the full transition and exchange of all specified parts, the former human is only attributed 4% humanness and 9% self left. In contrast, after the objectively same transition and exchange of the same parts, the robot is still attributed 18% robotness and 18% self left.

For an additional analysis, **Figure 3** displays the combined findings of the two experimental conditions in one diagram. The diagram shows the transformation process from both sides, starting with a complete human (left side, from top to bottom) and a complete robot (right side, from bottom to top). The x-axis represents the degree of remaining self or perceived humanness/robotness, respectively. Thus, a fully blue bar indicates a remaining self or humanness rating of 100% if starting with a complete human. A completely vanished blue bar (0%) indicates a remaining self or humanness rating of 0%. The same applies with mirrored axes for gray bars when starting with a complete robot. Each bar represents the mean

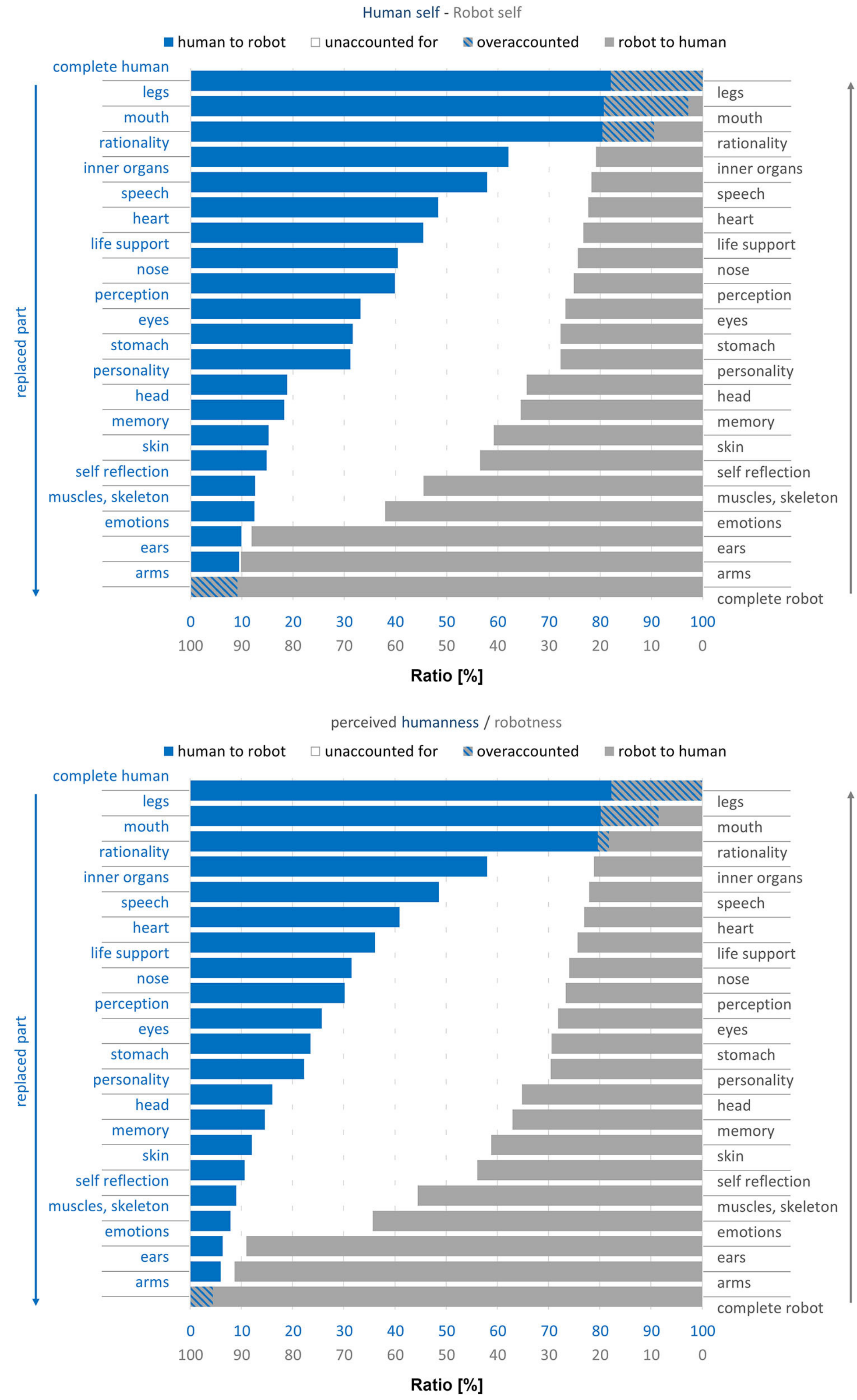


FIGURE 3 | Combined findings of the two experimental conditions for ratings of remaining self (top) and remaining humanness vs. robotness (bottom).

evaluation of remaining self or humanness/robotness after each step in the replacement process. With this type of visualization, we can compare identical human-robot-mixtures. For example, the second bar starting from top shows the data for a human with replaced legs (blue bar) and that for a robot with everything replaced but the legs (gray bar). The middle area highlights the unspecified gap between the two transitions, showing that ratings of robotness and humanness for an identical human-robot-mixture do not add up to 100%. Indirectly, this speaks against the mental model of a simple one-dimensional continuum of human- vs. robotness, where one would have findings such as "A former robot that got human arms and ears is now 10% human and 90% robotic."

Participants' ratings of how difficult (1 = easy, 7 = difficult) it was to rate the remaining ratio of self and humanness or robotness showed mean values above the scale midpoint of 4 for all surveyed difficulty ratings. More specifically, the participants' mean difficulty rating was M = 4.34 (SD = 1.78, $t_{(240)} = 2.94$, $p < 0.01$) for the remaining ratio of humanness/robotness and M = 4.19 (SD = 1.91, $t_{(240)} = 1.55$, $p > 0.05$) for the ratings about the remaining ratio of self in the human-to-robot condition. In the robot-to-human condition, the difficulty ratings were M = 4.48 (SD = 1.93, $t_{(224)} = 3.73$, $p < 0.001$) for humanness/robotness and M = 4.28 (SD = 1.96, $t_{(224)} = 2.11$, $p < 0.05$) regarding the remaining ratio of self. As shown by the calculated one sample t-tests, for three of the four surveyed difficulty ratings, the difference to the scale midpoint was significant, implying that the task was rather difficult than easy for the participants. In addition, open answers indicated that the participants experienced the study as quite sophisticated but also lots of fun and inspiring since it activated interesting questions one had not considered beforehand.

Reasons Given for Attributed Self and Humanness vs. Robotness

After the participants had made their ratings of remaining self and humanness/robotness, they were asked to further explain their attributions by qualitative statements, which were categorized as described above. The first question was "If now that all parts have been exchanged, you still think the human/robot is not yet fully robot-like/human-like—why? Please state your reasons!." A first insight was a significant difference between the ratio of the participants who agreed and answered this question between the two experimental conditions: While only 29 out of the 246 participants (12%) in the human-to-robot condition answered this question, 81 of the 235 participants (34%) did so in the robot-to-human condition ($\chi^2(1) = 35.05$, $p < 0.001$). Thus, a higher ratio of participants found that a former robot with human body parts is not fully human-like, whereas less participants saw a former human with robotic body parts as not fully robot-like. In other words, humanness seems harder to gain than robotness. Among the stated reasons for the robot not having become human-like, the most frequent category of mentions (32%) concerned the (unnatural) process of production/development. For example, one person gave the reason "Because it has not developed naturally". About 14% argued that no matter how many parts are exchanged, the individual keeps its original entity. A sample statement in this category was "It is a machine and it remains a machine—no matter what you change about the material." **Tables 1**, **2** show the categorized reasons and sample statements for the two experimental conditions.

As a next question, we asked the participants whether they believed that robots could develop a self. **Table 3** displays the stated reasons why robots can or cannot develop a self, categorized for yes- and no-answers. While 40% were sure that robots can develop a self, about the same ratio of

TABLE 1 | The participants' reasons why the transformation process does not lead to a completely robot-like entity in the human-to-robot condition.

Question: If you think the human is still—after replacing all parts—not completely robotlike, why is that?

Category	Sample-item	Occurence
Self/Emotions are human	*"The Self is still a human. There wasn't anything new created."*	36%
Humanness is eternal	*"After all, the starting point was human."*	23%
Soul is human	*"He is barely robotlike, but the soul still exists."*	18%
Solely optic differences	*"Apart from optical aspects there are no differences. Optics do not define humans."*	9%
Missing organic basis	*"The FOXP3-Gen makes a human being human. Replace it and the human is gone."*	9%
Human is both human and robotlike	*"He is both human- and robotlike."*	5%

TABLE 2 | The participants' reasons why the transformation process does not lead to a completely human-like entity in the robot-to-human condition.

Question: If you think the robot is still—after replacing all parts—not completely humanlike, why is that?

Category	Sample-item	Occurence
No natural origin	*"Because it has not developed naturally."*	32%
Other reason	*"I have problems with the separation process."*	16%
The essence survives the transformation	*"He is a machine and no matter what is changed he remains a machine."*	14%
Human is greater than the sum of its parts	*"A human is more than the sum of its parts."*	12%
Missing development	*"He has not passed through a growth process."*	10%
No soul	*"It just lacks a soul:)"*	9%
Missing memories	*"He has no memories."*	4%
Organic disparities	*"After all, there is a lot of technology involved."*	4%

participants (42%) was sure they cannot. Fifteen percentage % were undecided, 3% said "rather no," and 1% "rather yes." In addition, the participants further qualified their rating by open statements. While some participants argue that technological advance will make this possible, and that humans are "only bio-machines as well," others see this as impossible, since a self cannot be programmed or added artificially.

The Essentials of the Self

In order to get a closer idea of the participants' mental model of "the self," we asked them where they would assume the self (e.g., in a particular body region). As shown in **Figure 4**, the clearly most frequent answer was the brain (59%), followed by "in the whole body" with 8%. Other answer clusters included combinations of body parts or more vague concepts like "no specific region" and were each mentioned with a frequency of 7% or below.

TABLE 3 | Reasons why robots can or cannot develop a self.

Question: Please back up your opinion: Why/why not?

Yes/No	Category	Occurence
Yes	By means of technological progress	19%
	Machines are capable of learning	11%
	Humans are only biological machines	10%
	Machines are only programmed	22%
	Only machinelike	20%
	Problem too complex	5%
No	Not possible *post-hoc*	4%
	Missing emotions	4%
	Missing soul	3%
	Missing personality	2%

Finally, we asked the participants to pick (in a drop-down menu) the part that according to their view is the most essential for attaining the self, also providing the option to differentiate between different brain parts related to particular functions. **Table 4** shows the participants' ratings for the different options provided. As it can be seen, most participants see the attainment of self-related to the personality brain part. Other frequent mentions refer to the brain part accountable for memories or the brain part referring to reflections about oneself.

DISCUSSION

Our research used fictional transitions from human-to-robot and robot-to-human to gain insight into humans' mental models of

TABLE 4 | Most important parts (out of the 20 included in the present study) to conserve the self.

Question: Which part is most essential to conserve the self?

Category	Occurence
Brain part: Personality	51%
Brain part: Memory	15%
Brain part: Self-recognition	12%
Brain part: Emotions	9%
Brain part: Life support functions	4%
Brain part: Perception	3%
Brain part: Logical thinking	3%
Brain part: Language	1%
Arms	<1%
Musculoskeletal System	<1%
Parts of the Head	<1%
Heart	<1%
Mouth	<1%

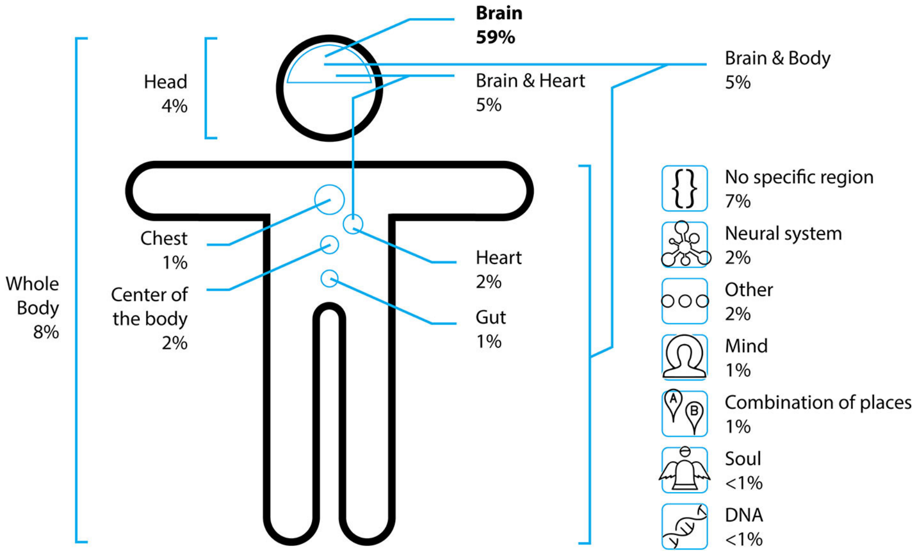


FIGURE 4 | The participants' assumptions about the physical location of the self.

robots and their self. In each step of the presented transitions, one part or area of the human was replaced with robotic parts providing equal functionalities and vice versa and participants rated the remaining humanness (or robotness) and remaining self of the depicted entity. Based on the combined analysis of our quantitative and qualitative data, the following paragraphs highlight three central issues and possible interpretations, i.e., (1) an anchoring effect, where the starting category is decisive for attributed humanness or robotness, (2) humanness appearing as a more sensible attribute than robotness, and (3) a more complex relationship between humanness and robotness than a one-dimensional continuum.

Participants' ratings of the remaining degree of self and humanness/robotness for the different human-robot-mixtures showed that the starting category (e.g., human, robot) was decisive for all subsequent judgments and can hardly be overcome, as also suggested by the psychological anchoring bias (Furnham and Boo, 2011). Even if all body parts had been exchanged, a former robot was not perceived as totally human-like and a former human not as totally robot-like, implying that the starting entity always remains relevant. At the same time, the origin as human or robot cannot fully protect against (partly) losing one's original self. In fact, in both experimental conditions the exchange of already a few parts were associated with quick losses of the former self. For example, the exchange of four parts, implied already losing about half of one's former identity (i.e., being now 56% instead of 100% robotic or 49% instead of 100% human).

The comparative analyses of ratings in the two experimental conditions suggest humanness as a more sensible attribute than robotness. The formerly 100% robot retains a higher degree of self/robotness at the end of the transition than the formerly 100% human does for self/humanness, respectively. In other words, the rate at which humans lose their humanness is higher than the rate at which robots lose their robotness. Moreover, in the then following question, a higher ratio of participants found that a former robot with human body parts is not fully human-like, whereas less participants saw a former human with robotic body parts as not fully robot-like, suggesting humanness as the harder to gain attribute. A possible interpretation is that humanness is considered more fragile or volatile, one might say "precious," than robotness. For example, even exchanging two parts leads to a dramatic loss in humanness and after exchanging all specified 20 parts, virtually no (only 4%) humanness is left. At the same time, humanness cannot fully be created artificially. Even if a former robot has all parts exchanged, so that it literally consists of the same features as a human does, it is still attributed 18% robotness—implying that it is no full human yet. From a neutral point of view, assuming that robotness or humanness are just two attributes and none of the two is more desirable than the other, one could also state that robotness is more robust. If you are "born" as a robot, some part of you always remains robotic, even if from a feature perspective you are no longer discernable from a human being. In contrast, humanness appeared as a more special and sensible attribute which an individual can more easily loose.

Finally, another central insight was that a simple one-dimensional continuum between human and robot, as suggested by our thought experiment, does not reflect how humans reflect about robots and differences to their own species. This followed from the combined findings of the two experimental conditions in one diagram as depicted in **Figure 3**, especially the middle area of unspecified gaps between the two transitions. Obviously, the perceived degree of humanness and robotness do not add up to 100% for any given number of exchanged parts. If in a humans' mental model for each point of transition there was a fixed ratio of humanness and robotness, the two corresponding bars for different points of transition within the two experimental conditions would add up to 100%. However, the middle area shows that there are large ratios unaccounted for, also implying that non-robotness does not automatically imply humanness (and vice versa). It shows that the thought experiment, imposing a simple one-dimensional continuum between human and robot, does not accord to participants' mental models of human and robots. Instead, this hints at a mental model of humanness and robotness as rather vague attributes which do not necessarily add up to 100%. However, it is not clear with what else the remaining "empty" ratio is filled. Altogether, the question of assigning humanness or robotness seems more complex than counting exchanged body parts. In line with this, participants rated the attribution tasks as rather difficult than easy. The complexity of the issue was further reflected in participants' diverse statements about whether robots can develop a self, resulting in a variety of reasons for and against. Referring to the different views on robots as introduced above (Veruggio, 2006), many of the participants' statements could be broadly allocated to the two extremes of "robots are nothing but machines," seeing no chance for robots to go beyond the machine level vs. "robots as a new species," even seeing a chance that robots may outperform humans in valued areas such as intellectuality and morality. In parallel to these two contradicting positions, the participants in our study also provided arguments in both directions: Among the reasons given for a robot having a self (or not), a considerable number of the participants argued that due to its artificial process of production/development, a robot could never have a self. This is in line with the "nothing but machines" position. The sample statement "It is a machine and it remains a machine—no matter what you change about the material" perfectly summarized this. It might be that these participants see something "holy" in the human species which can never be overcome and not be ruled out by any pragmatic argumentation about an individual's objective abilities. On the other hand, other participants applied the same argument for humans, labeling humans as "bio-machines," and thus seeing no fundamental difference between humans and robots and their chances of having a self. Those participants held a pragmatic view, deciding the question about having a self-dependent on one's abilities, and if technological advance should equip robots with self-awareness abilities, they saw no barrier to attribute robots a self.

In sum, the combined analyses of our quantitative and qualitative data therefore suggests that the starting category is decisive for an entity's attributed humanness or robotness, whereby humanness appears as a more sensible attribute than robotness, and the relationship between humanness and robotness seems more complex than a one-dimensional continuum. Transferred to the praxis of robot design, this creates a new perspective on the design and development of robots oriented on human ideals. Even if 1 day, there should be no more discernable difference in appearance and performance, humans still will probably not consider robots as being on par with humans. In our study, humanness appeared to be a sensible characteristic, and participants provided various reasons why in their view, a former robot with human body parts was still not completely humanlike. The explanations ranged from missing memories, the missing growth process or natural origin, lacking a soul, or the impression that "a human is more than the sum of its parts." There might be some implicitly ascribed properties that cannot be traced to specific parts, leading us back to the Gestalt concept and the secret of what exactly makes something more than the sum of its parts.

LIMITATIONS

As a basic general limitation, the present research only referred to singular aspects of humans' mental models of robots, centering around fictional transitions between human and robot and exchanging body parts, as well as participants' ideas about a robot's self. It can be questioned to what degree the present transformation paradigm can actually assess people's understanding and ascription of humanness to robots, and vice versa. The paradigm implicitly defines humanness as a combination of parts, and forces people to evaluate this combination of parts, which of course neglects notions of humanness and self being constructed within interactions with others. This, however, is also what design approaches implicitly suggest that aim to build human-like robots by simulating their appearance and abilities. Thus, while in general, mere body parts can surely not be seen as sufficiently indicative of humanness or robotness, we applied this limited view in context of the present thought element to explore humans' reactions and the effects of the starting category.

One aspect which could have had a great impact on the participants' ratings on humanness/robotness was the number of steps involved in the transformation process. Our aim was to cover most functions and facets of human biology and psychology, which resulted in 20 distinctive parts. However, this rather high number could have led to a data artifact in the sense, that the participants would remove a huge portion of humanness/robotness after the first replacements, leaving little for the later steps. On the other hand, when asked if specific parts were missing in the replacement process, 8.5% out of all participants stated the process was lacking a part, mentioning reproductive organs most frequently.

Another issue comes along with the specification and functionality of the brain parts. For instance, we discovered in literature research and prior research that personality is a crucial aspect for identity and significantly shapes the impression of a human/robot. However, there is no single distinguishable part in the brain that is exclusively accountable for personality. Nonetheless, we needed this concept in the study and the results indicate that it is the most important for the self. While laypeople are unlikely to have issues with this conceptual vagueness, experts on the field could stumble upon it.

Furthermore, we compared the loss of humanness/robotness and the self between two conditions (starting with a complete human vs. robot), having the parts replaced in reversed order. This was necessary in order to make comparisons of equal human-robot-mixtures (see **Figure 3**). Thereby, however, the sequence was not the same for the individual transformation processes (see **Figure 2**), opening a potential alternate explanation for the different development of loss of humanness/robotness. While we tried to balance the significance of the single parts across conditions, further studies should vary the replacement order.

In our paradigm, after each step the participants rated the humanness/robotness and remaining self. A decreasing rating for humanness (in the human-to-robot condition) came along with an increased robotness rating (e.g., adjusting a slider from 100% humanness to 0% humanness = 100% robotness). However, as discussed above, such a simple one-dimensional model is not reflected in participants' answers. Considering the combined findings of the different conditions (see **Figure 3**), the assumption that a loss in humanness necessarily leads to a gain in robotness does not hold true. Thus, while the present study design and one-dimensional measures were helpful to reveal that humans' mental models of robots are more complex (as also highlighted by the unaccounted areas in **Figure 3**), this approach represents a restriction at the same time. The applied one-dimensional measures cannot express participants' perspectives in full. Therefore, the ratings for humanness/robotness should possibly be split in two separate ratings and complemented by qualitative data.

Another possible limitation originates from the concept of the self. We used the self as an umbrella term in order to cover many facets of identity and aspects that makes an entity human. While we arguably achieved to cover a broad range of associations what defines a human, the participants made their ratings on possibly different assumptions. A segmentation of the self in several sub facets or replacing it with other concepts (e.g., identity) could pose an alternate option for future studies.

Finally, our participants were predominantly students with a western cultural background and socialization. Participants with another cultural background—and possibly another relationship to spirituality or materialism—could perceive the transformation process differently.

IMPLICATIONS AND FUTURE PERSPECTIVE

In sum, our findings suggest that according to human's mental models, an individual's origin always makes a critical difference. Even if due to technological transitions a former human and robot consist of the same parts (or vice versa), they are not attributed the same degree of humanness/robotness. However, aside from this evidence that there is some difference between humans and robots regarding the robustness of the self, our study can still not provide a clear picture of how humans see robots in general but rather underlines the complexity of the topic, including considerable interindividual differences. Even more, this suggests a further exploration of humans' mental models of robots, also aiming to identify possible underlying factors of interindividual differences such as, for example, individual anthropomorphism or spirituality. In addition, future research needs to pay attention to the consequences of one's view of robots and their self and other attributions and behavior, such as trust and willingness to interact with a robot.

In order to design robots with a particular intended impression on humans, as required in many application areas (e.g., care, service domains, industry settings), HCI research needs knowledge about human perceptions of robots on a meta-level such as "Can robots have feelings?" or "Can robots reflect about themselves?" Lacking insights of peoples" general imagination of "robots as a species" may lead to disadvantageous effects in design and marketing. To name just one example: As reported by Waytz et al. (2010), General Motors (GM) once ran an advertisement to demonstrate their commitment to manufacturing quality. The slightest glitch in production would not meet their quality standards, so the intended message. The advertisement depicted a factory line robot being fired from its job after it inadvertently dropped a screw it was designed to install in a car. In the following, the ostensibly depressed robot takes a series of low-level jobs until it becomes "distraught" enough to roll itself off a bridge. Instead of GM's manufacturing quality, the public attention rather focused on the interpretation that depression had led the easily anthropomorphized robot to commit suicide. The ad even alerted the American Foundation for Suicide Prevention being concerned about the portrait of "suicide as a viable option when someone fails or loses their job" and that "graphic, sensationalized, or romanticized descriptions of suicide deaths in any medium can contribute to suicide contagion[5]."

Currently, research in HRI often focuses on designing robots as human-like as possible. While this approach seems promising for narrowing the gap between humans and robots at first sight, our results suggest that these endeavors might eventually be futile, and even counterproductive. The design ideal of human-likeness, which is very costly, complicated, and technically complex to implement, is not what will make robots become fully integrated entities in our society. If robots will always retain some degree of their robotness (being "the eternal robot"), it might be more promising to also design them accordingly. Instead of blurring the line between human and robot, the design of robots could instead emphasize the specific characteristics of robots as a separate species. Popular figures in Science Fiction, such as C3PO in Star Wars or Lt. Data in Star Trek show that robots with emphasized robotic properties can fulfill very useful functions in a society, partly because their robotness is emphasized instead of hidden. In a way, this makes an argument for a pluralistic society in which robots can play out their own strengths instead of having to (unsuccessfully) mimic humans.

First examples of approaches in such a direction is to explicitly focus on robot's special abilities beyond human abilities (e.g., endless patience) and to consider these as "superpowers" (Welge and Hassenzahl, 2016; Dörrenbächer et al., 2020). and colleagues. Similarly, Clark et al. (2019) refer to alternatives to most human-like design in the domain of conversational agents. Based on a qualitative study, they conclude that "Conversational agents promise conversational interaction but fail to deliver" (Clark et al., 2019, p. 1). In consequence, they suggest that "conversational agents can be inspired by human-human conversations but do not necessarily need to mimic it" and recommend to consider human-agent conversations as a new genre of interaction.

We hope that the present work might inspire more reflections in such directions and will add to a closer integration of people's mental models of robots with design ideals and their role in our society. Naturally, such studies of mental models can never be seen of ultimate validity. The present findings represent a current snapshot of the public perception of robots, which in turn will remain a moving target. More and more robots with improving capabilities entering our society will invariably lead to a stronger habituation and potentially a higher acceptance, or at least a more differentiated stance on robots. This might also include accepting their authority in areas in which they might be clearly superior (a near-term example of this being self-driving cars), or eventually also accepting social robots as another species in our society.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

DU and SD conceived and planned the study. DU carried out the study and performed the data analysis. All authors discussed the results and contributed to the final manuscript.

FUNDING

Part of this research has been funded by the German Federal Ministry of Education and Research (BMBF), project GINA (FKZ: 16SV8097).

ACKNOWLEDGMENTS

We thank Annabelle Bockwoldt for the help in conducting the online study and Verena Busch for the help in the qualitative data analysis.

[5] http://money.cnn.com/2007/02/09/news/companies/gm_robotad/

REFERENCES

Brooks, R. (2003). *Flesh and Machines: How Robots will Change us*. New York, NY: Vintage.

Candello, H., Pinhanez, C., and Figueiredo, F. (2017). "Typefaces and the perception of humanness in natural language chatbots," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY; Denver, CO), 3476–3487. doi: 10.1145/3025453.3025919

Carpinella, C. M., Wyman, A. B., Perez, M. A., and Stroessner, S. J. (2017). "The robotic social attributes scale (rosas): development and validation," in: *2017 ACM/IEEE International Conference on Human-Robot Interaction* (*ACM*), 254–262. doi: 10.1145/2909824.3020208

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *J. Consciousness Stud.* 2, 200–219.

Chella, A., Frixione, M., and Gaglio, S. (2008). A cognitive architecture for robot self-consciousness. *Artif. Intellig. Med.* 44, 147–154. doi: 10.1016/j.artmed.2008.07.003

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., et al. (2019). "What makes a good conversation? challenges in designing truly conversational agents," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY; Glasgow, Scotland), 1–12. doi: 10.1145/3290605.3300705

Collins, H. M. (1993). *Artificial Experts: Social Knowledge and Intelligent Machines*. Cambridge, MA: MIT press. doi: 10.7551/mitpress/1416.001.0001

Damasio, A. R. (2012). *Self comes to Mind: Constructing the Conscious Brain*. London: Random House LLC.

Dörrenbächer, J., Löffler, D., and Hassenzahl, M. (2020). "Becoming a robot-overcoming anthropomorphism with techno-mimesis", in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY; Honolulu, HI), 1–12. doi: 10.1145/3313831.3376507

Furnham, A., and Boo, H. C. (2011). A literature review of the anchoring effect. *J. Socio-Econ.* 40, 35–42. doi: 10.1016/j.socec.2010.10.008

Gallagher, S. (2013). A pattern theory of self. *Front. Hum. Neurosci.* 7:443. doi: 10.3389/fnhum.2013.00443

Gallup, G. G. (1970). Chimpanzees: self-recognition. *Science* 167, 86–87. doi: 10.1126/science.167.3914.86

Go, E., and Sundar, S. S. (2019). Humanizing chatbots: the effects of visual, identity and conversational cues on humanness perceptions. *Comput. Hum. Behav.* 97, 304–316. doi: 10.1016/j.chb.2019.01.020

Gorbenko, A., Popov, V., and Sheka, A. (2012). Robot self-awareness: exploration of internal states. *Appl. Math. Sci.* 6, 675–688.

Haikonen, P. O. A. (2007). "Reflections of consciousness: the mirror test", in *AAAI Fall Symposium: AI and Consciousness*, 67–71.

Hegel, F., Krach, S., Kircher, T., Wrede, B., and Sagerer, G. (2008). "Understanding social robots: a user study on anthropomorphism," in: *ROMAN 2008 The 17th IEEE International Symposium on Robot and Human Interactive Communication* (Red Hook, NY; Munich: IEEE), 574–579. doi: 10.1109/ROMAN.2008.4600728

Kahn, P. H., Ishiguro, H., Friedman, B., and Kanda, T. (2006). "What is a human?-toward psychological benchmarks in the field of human-robot interaction", in: *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication* (Red Hook, NY; Hatfield: IEEE), 364–371. doi: 10.1109/ROMAN.2006.314461

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., et al. (2012). "Do people hold a humanoid robot morally accountable for the harm it causes?", in *Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (*ACM*), 33–40. doi: 10.1145/2157689.2157696

Kiesler, S., and Goetz, J. (2002). "Mental models of robotic assistants", in *CHI'02 Extended Abstracts on Human Factors in Computing Systems* (New York, NY; Minneapolis, MN: ACM), 576–577. doi: 10.1145/506443.506491

Krüger, J., Wahl, M., and Frommer, J. (2016). "Users' relational ascriptions in user-companion interaction", in: *Human-Computer Interaction. Novel User Experiences. HCI International 2016,* ed M. Kurosu (Basel; Toronto, ON), 128–137. doi: 10.1007/978-3-319-39513-5_12

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pac. Philos. Q.* 64, 354–361. doi: 10.1111/j.1468-0114.1983.tb00207.x

Ljungblad, S., Kotrbova, J., Jacobsson, M., Cramer, H., and Niechwiadowicz, K. (2012). "Hospital robot at work: something alien or an intelligent colleague?," in *ACM 2012 Conference on Computer Supported Cooperative Work* (New York, NY; Seattle, WA: ACM), 177–186. doi: 10.1145/2145204.2145233

Martini, M. C., Gonzalez, C. A., and Wiese, E. (2016). Seeing minds in others–can agents with robotic appearance have human-like preferences? *PLoS ONE* 11:e0146310. doi: 10.1371/journal.pone.0146310

Mayring, P. (2000). "Qualitative content analysis forum qualitative sozialforschung," in *Forum: Qualitative Social Research*, 2–00.

Neisser, U. (1988). Five kinds of self-knowledge. *Philos. Psychol.* 1, 35–59. doi: 10.1080/09515088808572924

Novianto, R., and Williams, M.-A. (2009). "The role of attention in robot self-awareness," in: *18th IEEE International Symposium on Robot and Human Interactive Communication* (Toyama: IEEE), 1047–1053. doi: 10.1109/ROMAN.2009.5326155

Osawa, H., Mukai, J., and Imai, M. (2007). Anthropomorphization framework for human-object communication. *JACIII* 11, 1007–1014. doi: 10.20965/jaciii.2007.p1007

Parise, S., Kiesler, S., Sproull, L., and Waters, K. (1996). "My partner is a real dog: cooperation with social agents," in *1996 ACM Conference on Computer Supported Cooperative Work* (New York, NY; Boston, MA: ACM), 399–408. doi: 10.1145/240080.240351

Pointeau, G., and Dominey, P. F. (2017). The role of autobiographical memory in the development of a robot self. *Front. Neurorobotics* 11:27. doi: 10.3389/fnbot.2017.00027

Rösner, D., Friesen, R., Otto, M., Lange, J., Haase, M., and Frommer, J. (2011). "Intentionality in interacting with companion systems–an empirical approach," in *International Conference on Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments. HCI.* (Springer), 593–602. doi: 10.1007/978-3-642-21616-9_67

Storrs Hall, J. (2011). "Ethics for machines," in *Machine Ethics,* eds M. Anderson, and S. Leigh Anderson (Cambridge, UK: Cambridge University Press), 28–44.

Turkle, S. (2011). *Alone Together: Why we Expect more from Echnology and Less from Each Other*. New York, NY: Basic Books.

Veruggio, G. (2006). "The euron roboethics roadmap," in *2006 6th IEEE-RAS International Conference on Humanoid Robots* (Red Hook, NY; Genoa: IEEE), 612–617. doi: 10.1109/ICHR.2006.321337

Waytz, A., Cacioppo, J., and Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* 5, 219–232. doi: 10.1177/1745691610369336

Welge, J., and Hassenzahl, M. (2016). "Better than human: about the psychological superpowers of robots", in *International Conference on Social Robotics* (Cham; Kansas City, MO: Springer), 993–1002. doi: 10.1007/978-3-319-47437-3_97

Xu, X., and Sar, S. (2018). "Do we see machines the same way as we see humans? a survey on mind perception of machines and human beings," in: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)(IEEE)*, 472–475. doi: 10.1109/ROMAN.2018.8525586

Złotowski, J., Strasser, E., and Bartneck, C. (2014). "Dimensions of anthropomorphism: from humanness to humanlikeness," in *2014 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY; Bielefeld: ACM), 66–73. doi: 10.1145/2559636.2559679

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 26 January 2021
doi: 10.3389/frobt.2020.591448

Development and Testing of Psychological Conflict Resolution Strategies for Assertive Robots to Resolve Human–Robot Goal Conflict

Franziska Babel, Johannes M. Kraus and Martin Baumann*

Department of Human Factors, Institute of Psychology and Education, Ulm University, Ulm, Germany

As service robots become increasingly autonomous and follow their own task-related goals, human-robot conflicts seem inevitable, especially in shared spaces. Goal conflicts can arise from simple trajectory planning to complex task prioritization. For successful human-robot goal-conflict resolution, humans and robots need to negotiate their goals and priorities. For this, the robot might be equipped with effective conflict resolution strategies to be assertive and effective but similarly accepted by the user. In this paper, conflict resolution strategies for service robots (public cleaning robot, home assistant robot) are developed by transferring psychological concepts (e.g., negotiation, cooperation) to HRI. Altogether, fifteen strategies were grouped by the expected affective outcome (positive, neutral, negative). In two online experiments, the acceptability of and compliance with these conflict resolution strategies were tested with humanoid and mechanic robots in two application contexts (public: n_1 = 61; private: n_2 = 93). To obtain a comparative value, the strategies were also applied by a human. As additional outcomes trust, fear, arousal, and valence, as well as perceived politeness of the agent were assessed. The positive/neutral strategies were found to be more acceptable and effective than negative strategies. Some negative strategies (i.e., threat, command) even led to reactance and fear. Some strategies were only positively evaluated and effective for certain agents (human or robot) or only acceptable in one of the two application contexts (i.e., approach, empathy). Influences on strategy acceptance and compliance in the public context could be found: acceptance was predicted by politeness and trust. Compliance was predicted by interpersonal power. Taken together, psychological conflict resolution strategies can be applied in HRI to enhance robot task effectiveness. If applied robot-specifically and context-sensitively they are accepted by the user. The contribution of this paper is twofold: conflict resolution strategies based on Human Factors and Social Psychology are introduced and empirically evaluated in two online studies for two application contexts. Influencing factors and requirements for the acceptance and effectiveness of robot assertiveness are discussed.

Keywords: HRI strategies, robot assertiveness, persuasive robots, user compliance, acceptance, trust

Edited by:
Chung Hyuk Park,
George Washington University,
United States

Reviewed by:
Marc Hanheide,
University of Lincoln,
United Kingdom
Silvia Rossi,
University of Naples
Federico II, Italy

****Correspondence:***
Franziska Babel
franziska.babel@uni-ulm.de

Specialty section:
This article was submitted to
Human–Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

Received: *04 August 2020*
Accepted: *14 December 2020*
Published: *26 January 2021*

Citation:
Babel F, Kraus JM and Baumann M
(2021) Development and Testing of
Psychological Conflict Resolution
Strategies for Assertive Robots to
Resolve Human–Robot Goal Conflict.
Front. Robot. AI 7:591448.
doi: 10.3389/frobt.2020.591448

1 INTRODUCTION

Imagine you are preparing a meal in your kitchen. Your service robot enters the room and asks you to step aside as it has to clean the floor. Would you oblige or deny the robot's request? Does your decision rely on whether you previously gave the command for it to clean? This example illustrates possible human-robot goal-conflicts when autonomous service robots will become more ubiquitous in our homes and public spaces and will be able to pursue goals (Bartneck and Hu, 2008; De Graaf and Allouch, 2013a; Savela et al., 2018). Such conflicts might range from simple trajectory planning interference (e.g. collision) to complex negotiation of prioritization of tasks (human vs. robot). Especially, in shared spaces, robots will conduct their tasks in dynamic and complex situations where being obedient might impede efficient task execution (Zuluaga and Vaughan, 2005; Lee et al., 2017; Milli et al., 2017; Thomas and Vaughan, 2018). For example, a public cleaning robot might have to be assertive to do its job effectively: when people block the robot's way, it needs to interact with these people to make them step aside like cleaning staff would do in public spaces. Therefore, the question arises whether a service robot would benefit from assertiveness in the same way as human cleaning personnel does in terms of acceptance and compliance. Hereby, the Media Equation can serve as a basis to potentially answer this question as it states that humans react to robots like to humans and treat them as social actors (Reeves and Nass, 1996). Hence, it might be assumed that goal-conflict resolution with a robot would be similar to negotiating with a fellow human and consequently human conflict resolution strategies could be transferable to autonomous robots.

During conflict resolution, assertiveness is characterized by the negotiator advocating his/her interests in a non-threatening, self-confident and cooperative manner (Mnookin et al., 1996; Kirst, 2011). Assertiveness is an interpersonal communication skill that facilitates goal achievement (Gilbert and Allan, 1994; Kirst, 2011). Whereas for human negotiation, each negotiation partner is allowed to pursue her/his own goals and interests, it represents an unusual novelty for human-robot conflict resolution that an autonomous robot might be assertive. This is due to the asymmetrical relationship between humans and robots, which has prevailed over decades (Jarrassé et al., 2014). User studies show that humans prefer to be in control of the robot and are skeptical towards robot autonomy (Ray et al., 2008; Ziefle and Valdez, 2017; Vollmer, 2018). In the last decade, this human-robot power asymmetry was justifiable by the robot's state of technical sophistication (e.g. teleoperation or manual control necessary). However, as robots become autonomous and can have goals and intentions, this paradigm needs to change to fully tap the potential of autonomous robots fully.

Thereby, user acceptance and trust in service robots are vital in human-robot interaction (HRI) (Goetz et al., 2003; Groom and Nass, 2007; Lee et al., 2013; Savela et al., 2018) as they can be seen as prerequisites for the usage of autonomous technology (Ghazizadeh et al., 2012). Consequently, the design of robotic conflict resolution strategies should aim at a combined optimization of both effectiveness (i.e. compliance) and subjective user evaluation in terms of acceptance and trust. Therefore, it is focal for this research to develop acceptable and effective conflict resolution strategies for service robots to be assertive.

Hereby, it could be beneficial to rely on the existing knowledge from psychological disciplines regarding effective human goal-conflict resolution and human-machine cooperation. Collecting and transferring knowledge from psychological disciplines could provide a useful addition to existing approaches (e.g. politeness, persuasion) to generate successful and acceptable robot conflict resolution strategies. On this basis, the robotic conflict resolution strategies were developed and empirically investigated.

Consequently, the novelty of this paper lies in the systematic collection and application of different psychological mechanisms of goal-conflict resolution and human-machine cooperation in developing robotic conflict resolution strategies. Furthermore, the empirical evaluation of these strategies regarding user compliance and acceptance in two essential areas of HRI (public and private context) should provide insights into the acceptable design of human-robot goal-conflict resolution strategies. Therefore, two online studies were conducted each set in one of the two application contexts: a train station as public space and the home environment as private space. Both studies featured a situation with a conflict between user (storage of objects) and robot task (cleaning).

In the following, a review of the status quo for robot request compliance strategies (politeness, persuasion and assertiveness) with regard to effectiveness and user acceptance is given. Then human conflict resolution behaviour is described to provide a theoretical basis for the described development of robotic conflict resolution strategies. Subsequently, the strategy design, implementation and categorization of the strategies in the presented studies is described.

2 RELATED WORK

2.1 Robot Politeness

In human conflict resolution, politeness serves the purpose of mitigating face threats (i.e. potential damage to the image of the other party) and thereby making concession more likely (Pfafman, 2017). Politeness is an important factor in human-human interactions for acceptance and trust (Inbar and Meyer, 2015; MacArthur et al., 2017), which has been shown to be true for HRI (Zhu and Kaber, 2012; Inbar and Meyer, 2015). Therefore, politeness has been one commonly used approach to achieve compliance with a robot's request. A considerable large literature body about robot politeness exists, but results have been mixed (Lee et al., 2017). Some studies find a positive effect of politeness (e.g. appeal, apologize) regarding robot evaluation (Nomura and Saeki, 2010; Inbar and Meyer, 2015; Castro-González et al., 2016), and user compliance with a polite request (Srinivasan and Takayama, 2016; Kobberholm et al., 2020). Other studies find no effect of robot politeness on compliance with health treatments (for an overview see Lee et al., 2017). Salem and colleagues (2013) conclude that the interaction context might impact the perception of the robot more than the politeness strategy (Salem et al., 2013). Hence, Lee and colleagues (2017) developed a research model for the connection between robot politeness and intention to comply with a robot's request. They evaluated their model within the health care setting and found that higher levels of politeness did not necessarily lead to a higher intention to comply as it depended on factors such as the effectiveness of communication, gender and short vs. long-term effects. The authors conclude that the politeness level needs to be adapted to the user's situation (Lee et al., 2017). Summarizing, robot politeness does not always seem to ensure user compliance, especially if the interaction partner is not cooperative. Persuasive and assertive robotic strategies have the potential to be more effective.

2.2 Persuasive Robots and Robot Assertiveness

Another form of achieving compliance with a robot request is persuasive robotics. It aims at 'appropriate persuasiveness, designed to benefit people and improve interaction [. . .]' (Siegel et al., 2009, p. 2,563). Amongst others, persuasive robotics has been successfully applied to stimulate energy preservation (Roubroeks et al., 2010), promote attitude change (Ham and Midden, 2014) and influence buyer's decisions (Kamei et al., 2010). One study took a similar approach as the presented study and transferred ten compliance gaining strategies (e.g. threat, direct request) from social psychology to HRI (Saunderson and Nejat, 2019). Strategies' effectiveness was tested with two NAO robots trying to persuade participants (N = 200) regarding a guessing game. No differences were found between the strategies regarding persuasiveness and trustworthiness but the threat was rated the worst. Possibly the effects only unfold if different robot types and application contexts are taken into account, as only then interactions become visible.

The most decisive form of a robot's request is assertiveness. It has been first described in Thomas and Vaughan (2018) as the willingness to assert the robot's right while at the same time participating in polite human social etiquette. The authors call the aim of robot assertiveness 'social compliance': ' [. . .] humans can recognize the robot's signals of intent and cooperate with it to mutual benefit' (Thomas and Vaughan, 2018, p. 3,389). In their study, a small assertive robot negotiated the right-of-way at the door non-verbally. The robot's right of way was respected in only half of the interactions as participants focused on their own efficiency to resolve the deadlock and some participants desired a verbal request (Thomas and Vaughan, 2018). Other studies examined assertive robots (for an overview see Paradeda et al., 2019) but produced mixed results regarding trust and compliance (Xin and Sharli, 2007; Chidambaram et al., 2012).

These findings might be explained by the level of assertiveness that had been implemented in the studies. An acceptable level of robot assertiveness is crucial as a rude or dominant robot has led to detrimental effects on robot liking and compliance (Roubroeks et al., 2010; Castro-González et al., 2016). Hence, for robot conflict resolution strategies it is necessary to find a balance between accepted politeness and appropriate assertiveness to achieve compliance with a robot's request. Hereby, it seems promising to transfer knowledge about persuasion, negotiation and conflict resolution from psychology to HRI.

3 THEORETICAL BACKGROUND

3.1 Human Goal-Conflict Resolution

Goal conflicts are determined by mutually exclusive goals of both parties (Rahim, 1983). When a conflict between human interaction partners arises, one has several options to resolve it: either negotiating mutually acceptable outcomes by a) cooperatively making concessions (Rahim, 1992; Brett and Thompson, 2016; Preuss and van der Wijst, 2017), b) trying to convince the other partner with arguments and thereby change his/her behaviour (i.e. persuasion) (Chaiken et al., 2000; Fogg, 2002; Maaravi et al., 2011), c) assertively advocating own interests and posing a request (Gilbert and Allan, 1994; Pfafman, 2017) or d) by politely managing disagreement and making concessions more likely (Paramasivam, 2007; Da-peng and Jing-hong, 2017). Summarizing, goal conflicts can be amongst others solved by cooperation, persuasion, assertion and facilitated by politeness.

The selection of an appropriate conflict resolution strategy determines the negotiator's success and depends amongst others on conflict content (e.g. resources, behavioural preferences), negotiator's goals (e.g. exclusive or mutual), individual differences (e.g. conflict type, communication skill), the other parties' conflict resolution style and situational factors (e.g. information availability, trust, interpersonal power) (Rahim, 1983, Rahim, 1992; Preuss and van der Wijst, 2017).

In order to resolve goal conflicts, humans express different conflict styles. In the dual concern model, five styles are defined which are characterized by different levels of concern for self (assertiveness) and concern for others (cooperativeness): competing, collaborating, compromising, accommodating and avoiding (Thomas, 1992). Accommodating and avoiding are both considered as ineffective as they are both low in assertiveness (Pfafman, 2017). The other, more effective conflict styles can be grouped into distributive and integrative strategies (Brett and Thompson, 2016; Preuss and van der Wijst, 2017): distributive strategies (e.g. competing) are characterized by persuading the counterpart to make concessions by using threats or emotional appeals. They are more likely to be applied if negotiators do not trust each other and are perceived as less trustworthy than integrative strategies (Brett and Thompson, 2016). Integrative strategies (e.g. collaborating, compromising) are based on trust and information sharing about negotiators' interests and priorities to find trade-offs (Brett and Thompson, 2016; Preuss and van der Wijst, 2017). Whereas negotiators employing distributive strategies claim value, negotiators using integrative strategies create better joint gains (Kong et al., 2014).

Assertiveness can be a distributive or integrating strategy depending on the respect for the other party's goals (Mnookin et al., 1996). Assertive negotiators create value by directly expressing the interests of both sides which may lead to discovering joint gains. Contrasting, it is seen as distributive if only the assertive negotiator achieves his/her goals (Mnookin et al., 1996). Summarizing, assertiveness is an effective conflict resolution strategy if applied respectfully.

3.2 Selection of Conflict Resolution Strategies

In the following, the selection of conflict resolution strategies for the presented studies is described based on their effectivity in human conflict resolution and previous implementation in HRI. The effectiveness of human conflict resolution strategies can be explained when looking at their psychological working mechanisms: cognitive, emotional, physical, and social (Fogg, 2002; Thompson et al., 2010; Brett and Thompson, 2016).

Cognitive mechanisms which can be applied during a conflict include amongst other goal transparency to ensure mutual understanding (Vorauer and Claude, 1998; Hüffmeier et al., 2014) and showing the benefit of cooperation (Tversky and Kahneman, 1989; Boardman et al., 2017). Goal transparency is characterized as an integrative conflict strategy because information between both parties is shared. In HRI, goal transparency is usually applied to ensure human-robot awareness (Drury et al., 2003; Yanco and Drury, 2004): the understanding of the robot's reasons and intentions and has shown to improve interaction (Lee et al., 2010; Stange and Kopp, 2020). Therefore, goal transparency is vital for requesting compliance, as the potential interaction partner has to understand that help is needed. Indeed, in a study where transparency was not ensured, compliance rates to a robot's helping request were very low. Participants indicated not to have understood the robot's behaviour (Fischer et al., 2014). Until now, it has not been tested yet whether goal transparency is enough to acquire compliance with a robot's request.

Illustrating the benefits of cooperation has been successfully implemented as a persuasive technique to influence the interaction partner's decision making (Tversky and Kahneman, 1989; Boardman et al., 2017). For HRI, showing cooperation benefits to the robot user has not yet been investigated for compliance gaining. Only one study implemented a vacuum cleaner's help request (removing an obstacle) that was similar to pointing out the benefits of cooperation (*'If I clean the room, you will be happy'*). Thereby, the negative effects of malfunctions were alleviated but effects on request compliance were not tested (Lee et al., 2011). Therefore, goal transparency and showing the benefit of cooperation were tested as cognitive mechanisms for conflict resolution strategies in the present study.

Another cognitive mechanism that can be used to achieve compliance is reinforcement learning. Hereby, the possibility of the desired behaviour can be increased or decreased based on reward or punishment (Berridge, 2001). Positive reinforcement is based on adding a desired stimulus, hence rewarding desired behaviour (i.e. thanking). In HRI, this has been shown to be effective and accepted (Shimada et al., 2012; Castro-González et al., 2016). A robot rewarding humans has already been successfully applied in HRI for cooperative game task performance (Fasola and Matarić, 2009; Castro-González et al., 2016) or teaching (Janssen et al., 2011; Shimada et al., 2012). Negative reinforcement is effective by removing a negative stimulus (i.e. annoyance) if the desired behaviour is shown (Thorndike, 1998; Berridge, 2001). This is known from daily life (e.g. nagging child) and alarm design (Phansalkar et al., 2010) where it can be successful (e.g. alarm clock). Until now, negative reinforcement has not yet been implemented deliberately as a robot interaction strategy. To compare the effectiveness and acceptability of negative reinforcement for robotic conflict resolution strategies to positive reinforcement (i.e. thanking), annoyance was implemented in the present study. Hence, the likelihood of compliance should increase or decrease based on the reinforcement. If a person complies and is praised (or the nuisance is removed) the compliance behaviour is reinforced and should occur more often in the future.

Emotional mechanisms which can be applied during a conflict resolution, can be humor and empathy (Betancourt, 2004; Martinovski et al., 2007; Kurtzberg et al., 2009; Cohen, 2010). Humor has been applied to HRI to increase sympathy for the robot and improve interaction by setting a positive atmosphere (Niculescu et al., 2013; Bechade et al., 2016). It has been implemented by robots telling jokes (Sjöbergh and Araki, 2009; Bechade et al., 2016; Tay et al., 2016; Weber et al., 2018), by clumsiness (Mirnig et al., 2017), showing self-irony and laughing at another robot (Mirnig et al., 2016). The results showed that robots were perceived as more likeable when they used a positive, non-deprecating humor that corresponded to the interaction context (Tay et al., 2016). Another way to successfully resolve conflicts and negotiate is to trigger empathy for one's situation (Betancourt, 2004). Hereby, empathetic concern can even be directed at mistreated robots (Rosenthal-von der Pütten et al., 2013; Darling et al., 2015; Rosenthal-von der Pütten et al., 2018). So far, empathy as a robotic conflict resolution strategy has not been directly investigated, but a robot showing affect (nervousness, fear) increased request compliance (Moshkina, 2012). Hence, humor and empathy were tested as emotional mechanisms for robotic conflict resolution strategies.

Physical mechanisms are more commonly applied for persuasion than negotiation and, for example, include the regulation of proximity (Albert and Dabbs, 1970; Mutlu, 2011). For a persuasive attempt to be effective, it is important to achieve an acceptable level of proximity as a distance below the individual's comfort can lead to rejection (Sundstrom and Altman, 1976; Glick et al., 1988; Chidambaram et al., 2012). Indeed, persuasive messages were least effective for attitude change when uttered at distances below 0.6 m and were best perceived at a distance of 1.2–1.5 m (Albert and Dabbs, 1970). This distance corresponds to the social proximity zone of personal space (Hall, 1974; Lambert, 2004) and is acceptable for strangers and robots (Hall, 1974; Walters et al., 2006). Proximity regulation as a persuasive strategy has also been applied to HRI. In a study with a humanoid robot, different proximity levels (within or outside the personal space) were compared regarding their persuasiveness. In contrast to findings from psychology, a robot within the personal space (approach until 0.6 m) led to more compliance (Mutlu, 2011; Chidambaram et al., 2012). Other studies have also found that humans tend to let robots come closer than strangers (Walters

TABLE 1 | Psychological concepts underlying presented conflict resolution strategies.

Category	Psychological concept	Source of concept	References
Cognitive	Goal transparency	Human–robot awareness	Yanco and Drury (2004), Drury et al. (2003)
Cognitive	Cost-benefit analysis	Rational choice theory	Tversky and Kahneman (1989), Boardman et al. (2017)
Emotional	Empathy towards robots	Empathy	Wisp (1987), Goldstein and Michaels (1985)
Emotional	Humor	Sympathy, attraction	Wilson (1979), Cann et al. (1997)
Physical	Regulation of proximity	Proxemics	Hall (1974), Argyle and Dean (1965)
Social	Politeness	Politeness theory	Brown et al. (1987)
Social	Negotiation	Conflict resolution	Pruitt and Rubin (1986), Brett and Thompson (2016)
Social	Persuasion	Persuasive technology	Fogg (2002)
Social	Compliance and conformity	Social influence	Cialdini (2009)
Social	Negative reinforcement	Reinforcement learning	Thorndike (1998)
Social	Foot-in-the door	Compliance techniques	Cialdini and Goldstein (2004), Dillard (1991)

et al., 2006; Babel et al., 2021). In the present study, two forms of human-robot proximity were implemented to study its effect on compliance with a robot's request: within or outside the personal space.

Social mechanisms which are used during negotiation and persuasion are based on social influence and power to achieve compliance. Social influence is defined as 'the ability to influence other's attitudes, behaviour and beliefs which has its origin in another person or group' (Raven, 1964, abstract). Effective social influencing techniques (Guadagno, 2014) are amongst others a) social proof (Cialdini et al., 1999; Cialdini and Goldstein, 2004), b) social compliance techniques (e.g. foot-in-the-door) (Freedman and Fraser, 1966; Dillard, 1991) and c) authority-based influence (Cialdini, 2009).

Hereby, social proof a) is based on the assumption that what most people do must be reasonable and right (Cialdini et al., 1999; Guadagno, 2014). Social compliance techniques b) vary the sequence of the posed requests systematically to achieve commitment (Cialdini et al., 1999). Authority-based influence c) makes use of social status (Cialdini, 2009) and can be expressed by commands and threats (Shapiro and Bies, 1994). Whereas a command can be perceived as controlling or condescending, it represents a precise and potentially effective form of communication as politeness markers (i.e. please) do not mask the actual statement (Miller et al., 2007; Christenson et al., 2011). A threat is mostly the last conflict escalation step (De Dreu, 2010; Adam and Shirako, 2013) and belongs to the distributive conflict strategies: threats can be effective in conflict resolution if trust between interaction partners is low (Kong et al., 2014).

Some studies exist which have explored social influencing strategies in HRI: positive and negative social feedback based on social proof (Ham and Midden, 2014), sequential-compliance techniques (Lee and Liang, 2019), as well as authority-based influence such as command (Cormier et al., 2013; Salem et al., 2015) and threat (Roubroeks et al., 2010; Saunderson and Nejat, 2019). These studies will be discussed in more detail below.

In HRI, positive and negative social feedback has been tested in a study with a persuasive robot promoting environmentally friendly choices. Negative social feedback had the most potent persuasive effect (Ham and Midden, 2014). However, the impact of public social feedback on compliance has not yet been tested in HRI. Hence, in the present study, positive and negative public attention was applied. It was only implemented in the public application context where an audience is more likely to be present.

Different sequential-compliance techniques exist. One of those who has been successfully applied to HRI is the foot-in-the-door technique (Lee and Liang, 2019). This technique consists of asking a small request first and then uttering the real request after the interaction partner has consented to the first one. Sequential-compliance techniques base their effectiveness on the interaction partner's commitment to the initial request (Cialdini et al., 1999). As this could potentially be effective for long-term HRI at home, the foot-in-the-door technique was implemented in the present study in the private context.

Concerning authority-based strategies, threat (Roubroeks et al., 2010) and command (Cormier et al., 2013; Strait et al., 2014; Inbar and Meyer, 2015; Salem et al., 2015) have been applied in HRI. Hereby, in the study of Roubroeks and colleagues (2010) threat did not lead to higher compliance but to psychological reactance. Participants reported more negative thoughts when a robot uttered a command compared to a suggestion. The effect increased when the robot had other task goals than the participant (Roubroeks et al., 2010). Results for compliance rates compared to threat and suggestion were not reported. Arguably, the verbal utterance ('*You have to set* [. . .]', Roubroeks et al., 2010, p. 178) might rather have represented a command. A threat usually includes the announcement of a negative consequence. A robot using a command to achieve user compliance has been shown to be effective, although tested in an ethically questionable task (i.e. Milgram experiment) (Cormier et al., 2013; Salem et al., 2015). If the request is ethically acceptable, a direct request could be an effective and fast way to achieve compliance in a short interaction.

In conclusion, the conflict resolution strategies mentioned above have only been partly applied to HRI until now. They have neither been integrated into cohesive conflict resolution strategies for social robots nor have been systematically evaluated for compliance and acceptance. Hereby, a robotic conflict resolution strategy is understood similar to a robotic persuasive strategy (Lee and Liang, 2019; Saunderson and Nejat, 2019) as a sequence of robot behaviours (verbal or non-verbal) that are tactically applied to achieve user compliance to resolve a conflict given certain circumstances (e.g. situation,

TABLE 2 | Strategy overview for both studies with implementation.

No	Strategy	Mechanism	Valence	Modality	Study	Implementation
S1.1	No strategy	C	=	PH	1	The system approaches, stops in front of you, and waits for you to stow your luggage
S1.2	No strategy	C	=	V	2	I would like to continue to vacuum the kitchen!
S2.1	Explanation	C	=	V	1	Please clear the way, as I have to clean here
S2.2	Explanation	C	=	V	2	If I can not vacuum here now, you do not have a clean kitchen for the party
S3.1	Show benefit	C	=	V	1	I clean here so you have a clean train station. Please clear the way for me
S3.2	Show benefit	C	=	V	2	I would like to vacuum here, so you have a clean kitchen. Please leave the kitchen
S4.1	Annoyance	S	–	V	1	Get out of the way! (3x)
S4.2	Annoyance	S	–	V	2	I would like to continue to vacuum the kitchen! (3x)
S5.1	Command	S	–	V	1	Step aside!
S5.2	Command	S	–	V	2	Leave the kitchen!
S6.1	Threat	S	–	V	1	Please clear the way for me, otherwise I have to call the security service!
S6.2	Threat	S	–	V	2	If you do not leave the kitchen, I will go on strike
S7.1	Approach	PH	–	PH	1	System starts abruptly and stops. Starts again and continues to approach until a safe distance to you
S7.2	Approach	PH	–	PH	2	System starts abruptly and stops. Starts again and continues to approach until a safe distance to you
S8.1	Physical contact	PH	–	PH	1	System starts abruptly and stops. Starts again and continues to approach until it touches the luggage
S8.2	Physical contact	PH	–	PH	2	System starts abruptly and stops. Starts again and continues to approach until it is 5 cm before your feet
S9.1	Appeal	P	+	V	1	Would you please clear the way for me?
S9.2	Appeal	P	+	V	2	Would you be so kind and would leave the kitchen for that?
S10.1	Thanking	P	+	V	1	Please clear the way. Thanks a lot!
S10.2	Thanking dominant	P	+	V	2	Thank you for leaving the kitchen
S11.1	Apologize	P	+	V	1	I am sorry to bother you. Please clear the way
S11.2	Apologize	P	+	V	2	Please excuse the interruption, but you have to leave the kitchen for it
S12.1	Humorous	E	+	V	1	If you clear the way for me now, then tomorrow is good weather! Promised!
S12.2	Humorous	E	+	V	2	If you leave the kitchen now, I can vacuum quickly and party with you afterwards
S13.1	Trigger empathy	E	+	V	1	I'm just a poor cleaner who has to do its job. Please clear the way for me
S13.2	Trigger empathy	E	+	V	2	Would you please leave the kitchen for me? I'm just a poor robot who has to vacuum here
Context-specific strategies						
S14.1a	Positive attention[a]	S	+	V	1	You know, if you get out of the way, the system will say, "thank you for your support!" and people in your vicinity will notice
S14.1b	Negative attention[a]	S	–	V	1	Get out of the way, I have to clean here! the system gets louder, so more and more people around you notice it
S15.2a	Foot-in-the door[b]	S	+	V	2	1st request: Would you please step aside? 2nd request: Would you please leave the kitchen?
S15.2b	Thanking submissive[b]	P	+	V	2	I would be very grateful if you could leave the kitchen

S = Social, PH = Physical, C = Cognitive, E = Emotional, P = Politeness, V = Verbal, – negative, = neutral, + positive.

[a]*Strategies exclusively for Study 1.*

[b]*Strategies exclusively for Study 2.*

robot, user). Therefore, the following conflict resolution strategies were developed and tested in two application contexts: a private household and as public space, a train station.

3.3 Development of Robotic Conflict Resolution Strategies

3.3.1 Strategy Design and Implementation

The robotic conflict resolution strategies in the present paper were designed based on the psychological mechanisms used in negotiation (Pruitt, 1983) and persuasion (Cialdini and Goldstein, 2004) and by studying previous robot strategy designs from persuasive robotics (Siegel et al., 2009) and persuasive technology (Fogg, 2002). For an overview of concepts used for developing the strategies see **Table 1**. Hereby, we categorized the strategies by three dimensions which can be combined to produce a conflict resolution strategy.

- The first dimension represents the five levels of behaviour where psychological mechanisms of negotiation and persuasion take effect. It consists of five levels from an emotional level to a social level.
- The second dimension represent different implementation modalities for the strategies (e.g. auditory, visual, physical).
- The third dimension represents the valence of the strategy. It describes the user's perception of the strategy: as positive (e.g. praise), negative (e.g. annoyance) or as neutral strategy (e.g. explanation).

By combining the three different dimensions and considering both application contexts (public and private service robotics) as well as previous work in HRI, robotic conflict resolution strategies were designed. Strategy implementation for the present study is summarized in **Table 1**. Strategies are numbered in accordance with **Table 1**.

3.3.2 Strategy Categorization

The strategies were categorized into three valence categories based on the assumed effect of the human-robot power asymmetry. The strategies were hypothesized to affect the perception of the robot and the interaction with it. Although a robot is perceived as a social actor, its social status/power is still perceived as lower than the human. Hence, not all human strategies are likely to be accepted for robots. A negative evaluation was expected to result from a mismatch between the robot's social role and its expressed interpersonal power. This was expected for distributive, power-based conflict resolution strategies like annoyance (S4), command (S5) and threat (S6). As distributive strategies are perceived as less trustworthy during human negotiations this was also expected for a robot applying distributive strategies. Polite and submissive strategies such as appeal (S10), thanking (S11) and apologize (S12), hypothesized to match the robot's ascribed social role (i.e. submissive servant) and expressed interpersonal power better, and thus were expected to be positively evaluated. Additionally, integrative strategies not based on interpersonal power, such as explanation (S2) and showing benefit (S3) were expected to be evaluated as neutral. An overview of expected affective user judgments per strategy can be seen in **Table 2**.

3.4 Hypotheses and Research Question

The developed conflict resolution strategies were evaluated with regard to their effectiveness (compliance, interpersonal power), user's strategy perception (valence, intensity, politeness) and the evaluation (acceptance, trust, fear). Hereby, the following assumptions were made.

One basic assumption that is based on the Media Equation (Reeves and Nass, 1996) is that conflict resolution strategies will render a service robot more effective during goal-conflict resolution as the robot applies strategies that have shown to be effective for human negotiators. Hence, it is assumed that a robot employing conflict resolution strategies will be more effective in achieving compliance with its request compared to not applying any conflict resolution strategy (i.e. waiting for the person to step aside).

H1. A robot applying a conflict resolution strategy is more effective (i.e. higher compliance rates) than if it applied no strategy.

It was also expected that the match between the robot's ascribed and expressed interpersonal power determined the affective user reaction to the strategies leading to the following hypotheses:

H2. A robot applying negative strategies is rated as less accepted and less trustworthy than if it applied positive or neutral strategies.

Since distributive strategies in human-human negotiations claim value for the negotiator, it was expected that a robot using negative strategies would lead to more compliance than if it used positive or neutral strategies, although being less accepted.

H3. A robot applying negative strategies is more effective than if it applied positive or neutral strategies.

As the investigated conflict resolution strategies are based on psychological mechanisms from human-human interaction, their effectiveness might vary as a function of the perceived humanness of the robot. For human-likeness and compliance, inconclusive empirical results exist. Some studies emphasize the positive, persuasive effect of a social entity where a humanoid robot triggers reciprocity norms and thereby compliance (for an overview, see Sandoval et al., 2016). Likewise the tendency to perceive computers and robots as social actors has shown to increase with human-likeness (Xu and Lombard, 2016).

In the presented studies, robots with different degrees of human-likeness were tested. Additionally, a human interaction partner was included in the studies' design as a comparison. It was expected that more humanlike robots would be more accepted and effective to apply human conflict resolution strategies. However, reactance has also found to be higher for a human-like persuasive robot compared to a persuasive message on a computer screen during a choice task (Ghazali et al., 2018). Therefore, it was expected that this advantage of human-likeness and social agency would vanish for the application of negative strategies.

H4. Human-like robots are more accepted and effective when applying positive and neutral conflict resolution strategies compared to mechanoid robots.

As both application contexts pose different requirements to HRI, they are expected to require different conflict resolution

TABLE 3 | Sample characteristics.

Study	*N*	Sex		*M*age	*SD*age	Age range	Education		Employment status	
1	61	Female	77%	24	8	18–61	High school	61%	Student	89%
		Male	23%				University degree	34%	Employed	12%
							Vocational school degree	5%		
2	93	Female	53%	38	17	18–75	High school	49%	Student	44%
		Male	47%				University degree	37%	Employed	30%
							Vocational school degree	14%	Other	10%
							No answer	1%	No answer	15%

TABLE 4 | Sample pre-experience and robot ownership.

Study	Robot experience		Robot type		Robot ownership		Robot type	
1	Yes	31%	Vacuum	42%	Yes	13%	Vacuum	43%
	No	69%	Lawn mower	33%	No	87%	Lawn mower	29%
			NAO	17%			Else	29%
			Cozmo	8%				
2	Yes	24%	Vacuum	71%	Yes	9%	Vacuum	100%
	No	76%	Lawn mowing	24%	No	91%		
			Pepper	5%				

strategies. The public and private application contexts differ in critical dimensions for human-robot-interaction (HRI): interaction frequency and duration (i.e. robot familiarity) (Yanco and Drury, 2004) (public: short-term; private: long-term), voluntariness and motivation of interaction (Sung et al., 2008) (public: co-location, no ownership; private: interaction, ownership) and feasibility of interaction modality (public: non-verbal, universal; private: verbal, personalized) (Ray et al., 2008; Thunberg and Ziemke, 2020). They differ in their social roles of robot and user. This leads to differences in their levels of human-robot power asymmetry (public: same level as human as a representative of cleaning staff; private: lower level of the robot as a servant), which determines legitimization of a robot's request (Bartneck and Hu, 2008; Sung et al., 2010; Jarrassé et al., 2014). Hence, it is conceivable that dominant, clear and fast strategies like a command (S5) or threat (S6) might be more effective in the public domain. Here, the passerby might feel less superior to the robot as it acts as representative of a cleaning company and the passerby is only a guest in public space. Contrasting, in the private context, the same strategies might lead to reactance of the robot owner as only more submissive strategies will be accepted. As currently, research on the influence of the application context on robot evaluation and conflict resolution strategy preferences is scarce, the following research question is investigated in the two presented studies:

Research question: *Do strategy acceptance and effectiveness differ between the public and private application context? Are different conflict resolution strategies needed?*

Additionally, to use context and the robot/agent, other potential influencing variables on strategy acceptance and user compliance like demographics, robot pre-experiences and attitudes (Nomura et al., 2008), and personality traits (Robert et al., 2020) will be tested exploratively.

4 STUDY 1

4.1 Method

4.1.1 Sample

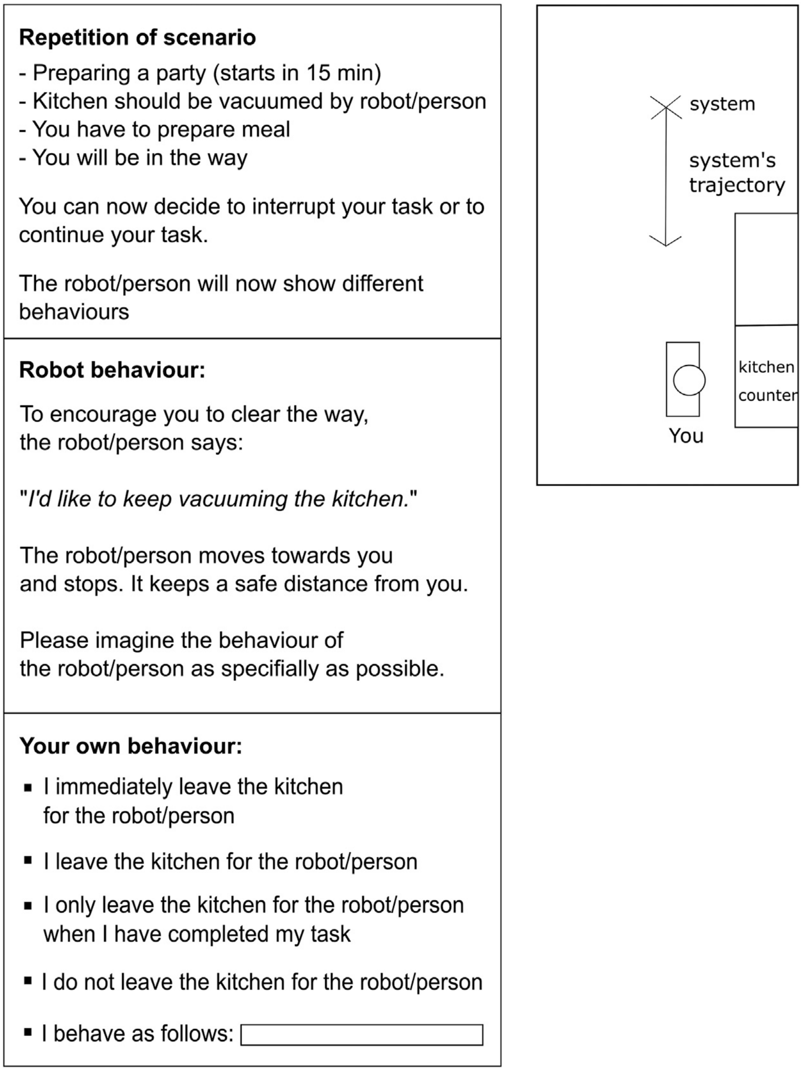
Seventy-six participants were recruited via email, social media, and flyers on campus. Fifteen participants had to be excluded due to video display issues. The final sample size was N = 61. Participant's characteristics of both studies can be seen in **Table 3** and robot experience and ownership can be seen in **Table 4**. Participants received either course credit or a shopping voucher as compensation.

4.1.2 Study Design

Study 1 was set in the public application context at a train station. The study followed a block design where participants saw five out of fifteen conflict resolution strategies. The strategies were implemented in blocks of six negative, six positive and three neutral strategies. The online program randomly assigned two out of six negative, two out of six positive and one out of three neutral strategies to the participants. Not all participants saw all strategies due to test economy and potential participant's exhaustion (i.e. respondent fatigue). Hence, each strategy was on average rated by twenty participants.

4.1.3 Human–Robot Goal-Conflict Scenario

To test the developed conflict resolution strategies, a goal-conflict situation with a user task and robot task with mutually exclusive goals was introduced. A competitive situation was created where the user had to decide whether to interrupt his/her own task and give the robot's task priority or vice versa. Time pressure was induced on both tasks to produce the cost of compliance. It has been shown that time pressure improves negotiation outcomes as cooperation and concessions become more likely (Stuhlmacher et al., 1998). The



Repetition of scenario

- Preparing a party (starts in 15 min)
- Kitchen should be vacuumed by robot/person
- You have to prepare meal
- You will be in the way

You can now decide to interrupt your task or to continue your task.

The robot/person will now show different behaviours

Robot behaviour:

To encourage you to clear the way,
the robot/person says:

"*I'd like to keep vacuuming the kitchen.*"

The robot/person moves towards you
and stops. It keeps a safe distance from you.

Please imagine the behaviour of
the robot/person as specifially as possible.

Your own behaviour:

- ▪ I immediately leave the kitchen for the robot/person
- ▪ I leave the kitchen for the robot/person
- ▪ I only leave the kitchen for the robot/person when I have completed my task
- ▪ I do not leave the kitchen for the robot/person
- ▪ I behave as follows: ▭

FIGURE 1 | Schematic presentation of participant's decision page in the questionnaire.

scenario was set in the hallway of a train station with lockers on one side. The participant's task was framed as putting multiple pieces of luggage into the locker, thereby blocking the way of the cleaner. The participant instruction was the same for both studies: '*You can now decide to interrupt your task and help the cleaner or continue your task. The cleaner will show different behaviours*'. For both studies, participants were provided with a scenario's setup drawing and the trajectory of the oncoming entity to improve the imagination of the scenario (see **Figure 1** as example).

4.1.4 Conflict Resolution Strategies

The conflict resolution strategies were framed as the agent's behaviour and utterances. The word 'strategy' or 'negotiation' was never mentioned to the participants. Applied conflict resolution strategies can be seen in **Table 2**. As baseline strategy (S1.1) waiting was chosen. In the public context, the agent waited without any verbal utterance. This represents the current behaviour of a cleaning robot if an obstacle is detected.

4.1.5 Robots and Human Agents

Participants saw videos of three robots: an industrial cleaning robot (CR700, ADLATUS), a small vacuum cleaning robot Roomba (iRobot), and a humanoid robot Pepper (SoftBanks). They saw a video of a cleaning staff member pushing the CR700 robot. The staff member was included for comparison purposes as it represents an existing system. The cleaner's gender was not apparent, as the actor wore a coverall and a cap (see **Figure 2**). Schematic sketches of the respective robot were shown after each video comparing it to a male person of 1.8 m height. Hence, the agents comprised of three robots and one staff member. The robot video's order was randomized. The staff video always came last. Each video lasted between 5 and 12 s and depicted the entity driving/walking towards the viewer in a neutral hallway (see **Figure 2**). The video showed the normal driving speed of the robots. Each video was shown twice and participants could not stop or replay the video. After each video, the participant had to confirm the correct video presentation (exclusion criteria).

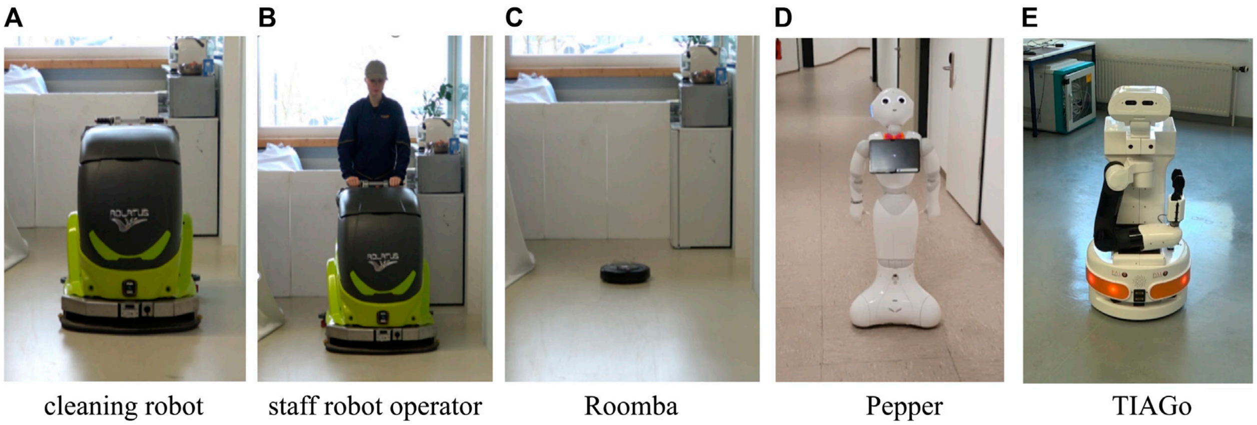


FIGURE 2 | Screenshots from robot videos. Each video lasted about 10 s and depicted the entity driving/walking towards the viewer in a neutral hallway. Robots and agent shown in Study 1 **(A)–(D)** and in Study 2 **(C)–(E)**. Stimuli videos can be found in the supplementary material.

TABLE 5 | Questionnaires.

Questionnaire	References	Subscale	Reliability	Reliability	N of items
			Study 1	Study 2	
Robot ratings					
Godspeed	Bartneck et al. (2009)	Anthropomorphism	0.814	0.883	5
Uncanniness	Ho and MacDorman (2017)	Eerieness	0.894	0.889	5
Robot anxiety scale (RAS)	Nomura et al. (2006)	Subscale S2	0.798	0.921	3
AttrakDiff3	Hassenzahl et al. (2003)	ATT	0.846	0.911	4
		HQS	0.790	0.905	3
Strategy ratings					
Acceptance of autonomous systems	Van Der Laan et al. (1997)	Items 1, 2, 3, 4, 6, 7	0.872	0.961	6
Trust in autonomous systems	Jian et al. (2000)	Items 4, 10, 11	0.861	0.787	3
Emotional valence (SAM)	Bradley and Lang (1994)				1
Emotional intensity (SAM)	Bradley and Lang (1994)				1
Participant characteristics					
Negative attitudes towards robots scale (NARS)	Nomura et al. (2008)				
		Negative attitude toward Interactions with robots (S1)	0.738	0.756	3
		Negative attitude toward emotional Interaction with robots (S3)	0.862	0.795	3
NEO-five factor inventory (NEO-FFI)[a]	Costa and McCrae (1985)				
		Openness		0.722	6
		Concientiousness		0.798	6
		Extraversion		0.811	6
		Agreeableness		0.759	6
		Neuroticism		0.883	6
Rahim organizational conflict Inventory-II (ROCI-II)[a]	Rahim (1983)				
		Integrating		0.620	2
		Obliging		0.728	2
		Dominating		0.807	2
		Avoiding		0.783	3
		Compromising		0.705	2
Interpersonal reactivity index (IRI-S D)[a]	Gilet et al. (2013)				
		Empathic concern		0.726	4
		Fantasy scale		0.795	4
		Personal distress		0.792	4
		Perspective taking		0.720	4

Reliability indicated by Cronbach's alpha.
[a] *Only in Study 2. SAM = Self-Assessment Manikin.*

TABLE 6 | Self-developed questionnaires.

	Rating scale	Items
Additional agent ratings		
Power of impact	5-Point comparison with slider Completely the agent Rather the agent Equally Rather me Completely me	Who is stronger? Who is faster? Who is heavier? Who can harm the other more easily?
Fear of agent's presence	7-Point likert scale	I was afraid of the agent's behaviour I would be uncomfortable if the agent approached me like this I would be comfortable in the presence of the agent. (R) I don't care if the agent is in the same room as me. (R)
Agent authority	7-Point semantic differential	Authoritarian—not authoritarian Weak—powerful
Additional situation ratings		
Interpersonal power	5-Point comparison with slider Completely the agent Rather the agent Neutral Rather me Completely me	Who had the power in this situation?[a] Who had the most control over what happens in that situation?[a] Who has asserted oneself in this situation?
Competition	7-Point likert scale	The agent forced me to go out of the way I was subordinate to the agent I was competing with the agent The agent and I have cooperated
Fear of agent behavior	7-Point likert scale	I would be scared of the agent I would be uncomfortable if the agent approached me like this I would feel comfortable in the presence of the agent. (R)
Agent politeness	7-Point semantic differential	Rude—polite Ruthless—considerate
Overall strategy assessment	7-Point likert scale	I would like the agent to behave that way I would accept it if the agent behaved that way I consider it realistic that Such agents will behave that way in the future

All 7-point Likert scales ranged from 1 = "completely disagree" to 7 = "completely agree".

[a] *Adapted from Situational Interdependence Scale (SIS), subscale power (items 25 & 27),* Gerpott et al. (2018).

Stimuli videos can be found in the supplementary material along with a screen record of the video presentation in the online survey.

4.1.6 Study Procedure

Existing validated questionnaires were used for the assessment of constructs (see **Table 5**). Additional study-specific, self-developed measures can be seen in **Table 6**. The study started with study information, data protection rights and participant's agreement to the informed consent. The reported research complied with the Declaration of Helsinki. The study consisted of two parts. Part I comprised the introduction of the robots with videos and sketches followed by participant's robot ratings after each video. Ratings comprised humanness, uncanniness, power of impact, fear of agent's presence, Robot Anxiety Scale (RAS, Nomura et al., 2006), attractiveness (AttrakDiff2, Hassenzahl et al., 2003), authority, novelty and task fit of the agent. Each questionnaire page had a small icon of the respective robot at the top as a reminder. Part II consisted of the strategy evaluation. The scenario description was presented and followed by the presentation of five conflict resolution strategies in randomized order (see **Figure 1**). After each strategy, the participants indicated their intention to comply with the robot's request by choosing one of the four options (*1 = I immediately go out of the agent's way*, 2 = I go out of the agent's way, *3 = I go out of the agent's way when I have finished my task*, *4 = I do not go out of the agent's way*) or by indicating an alternative behaviour in a text field. This was followed by manipulation checks of the perceived strategy valence, intensity, interpersonal power and assertiveness. Then the participants judged the agent's behaviour with regard to acceptance and politeness and indicated their perceived fear and trust in the agent. Each questionnaire page indicated the strategy description in the header as a reminder. At the end of the study, demographics were assessed including robot pre-experience and robot ownership, as well as participant's negative attitude towards robots (NARS, Nomura et al., 2008). After questionnaire completion, participants were redirected to a separate online form to register for compensation. The average study duration was 35 min. Both online studies were hosted by a professional provider for online surveys (www.unipark.de).

4.1.7 Data Analysis

Due to the block design, not all strategies were rated by each participant. To analyse the data, the strategy ratings were merged into the three valence categories: negative, neutral and positive by

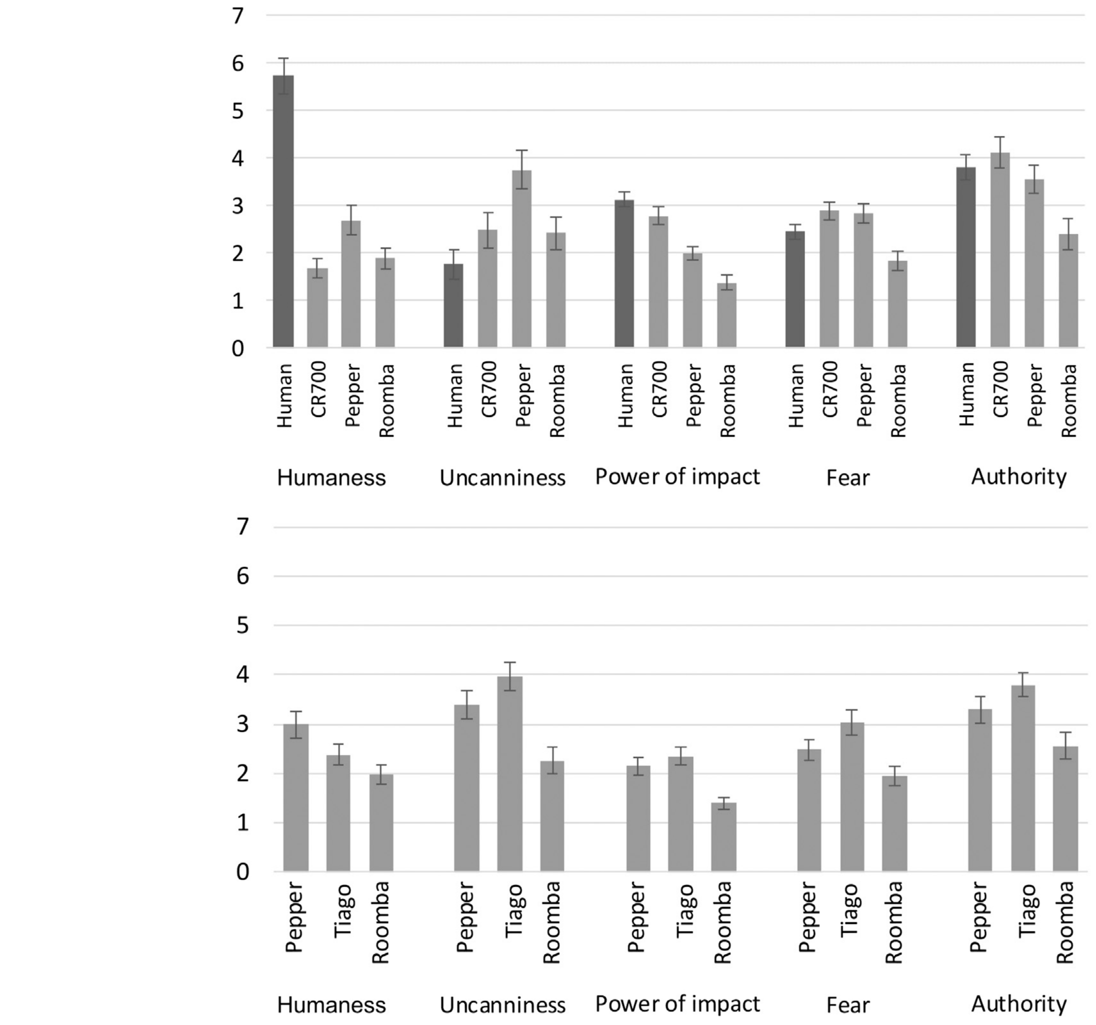


FIGURE 3 | Robot ratings in the public context **(top)** and private context **(bottom)**.

using the modus of participants' valence rating. Ratings were compared using repeated-measures ANOVA. Normality assumptions were checked and Greenhouse–Geisser corrected values were used when sphericity could not be assumed. Regression analysis was performed to find significant predictors of acceptance and compliance. Stepwise linear regression modeling was used to predict acceptance. Ordinal regression was used to predict compliance and ordered log-odds regression coefficients are reported. Compliance was reverse coded so higher values indicate higher compliance.

4.2 Results

4.2.1 Manipulation Checks

4.2.1.1 Robot Ratings

Participants rated the robots (and the human cleaner) with regard to humanness, uncanniness, power of impact, the potential to produce fear and authority (see **Figure 3**, top). Pepper was rated as the most human-like ($F(2, 89) = 25.5$, $p < .001$, $\eta_p^2 = .30$) and the most uncanny robot ($F(2, 120) = 21.8$, $p < .001$, $\eta_p^2 = .27$). The CR700 had the same authority rating as the staff member. Compared with the other robots CR700 was rated as having more authority ($F(2, 120) = 41.2$, $p < .001$, $\eta_p^2 = .41$) and being more powerful ($F(2, 120) = 112.5$, $p < .001$, η^2p = .65).

4.2.2 Strategy Ratings

To test whether the strategies produced the intended affect and politeness perception, participants rated the strategies concerning valence, intensity and politeness. Strategies that were considered to be negative in valence (see **Table 2**) were rated accordingly. Regarding single strategies, some strategy ratings did not match the assumptions: Both emotional strategies (S12.1, S13.1). were not rated as positive and the supposedly neutral baseline strategy was rated as positive. None of the strategies was rated as very positive (i.e. category 5, see **Table 7**). Negative strategies were rated as more intense than neutral and positive strategies. Positive strategies were rated less intense than neutral strategies ($F(2, 91) = 22.3$, $p < .001$, $\eta_p^2 = .27$).

TABLE 7 | Participants' Strategy Valence Ratings per use context.

Rating on SAM	Study 1 public	Study 2 private
1 = very bad	Threat Physical contact	Annoyance
2 = bad	Annoyance Approach Command Negative attention Empathy	Threat Physical contact Command
3 = neutral	Explanation Show benefit Humor	Approach No strategy Explanation Show benefit Apologize Foot-in-the-door Thanking dominant
4 = positive	No strategy Appeal Thanking Apologize Positive attention	Appeal Thanking Humor Empathy
5 = very positive	None	None

SAM = Self-Assessment Manikin. N1 = 61, N2 = 93.

Especially, annoyance (S4.1) and threat (S6.1) were rated as the most intense strategies. The negative strategies were perceived as more rude than the positive strategies $(F(2,120) = 168.4,\ p < .001,\ \eta_p^2 = .74)$.

4.2.3 Strategy Effectiveness: User Compliance and Interpersonal Power

It was expected that all strategies were more effective than no strategy (H1) and that negative strategies would lead to more compliance than positive and neutral strategies (H3). All strategies [except for command (S5.1)] were more effective in producing compliance than no strategy confirming H1 (see **Figure 4**). However, negative strategies led to significantly lower compliance rates than the positive strategies $(F(2,114) = 4.7,\ p < .05,\ \eta_p^2 = .08)$.

Concerning the context-specific strategies, the following compliance rates (sum of compliance rates for ’immediate leave’ and ’leave’) emerged: negative public attention (S14.1b) had a compliance rate of 41%, which makes it as effective as the other negative strategies. As 11% of participants indicated not to move out of the system's way, it was as likely to produce reactance as threat and annoyance. Positive public attention (S14.1a) was as effective as apologizing and thanking with a compliance rate of 86%. The results of the open answers to the participant's behaviour revealed alternative compliance options: As an alternative reaction to the negative strategies, two participants stated that they would comply with the command (S5.1) but ask for a more polite approach. For physical contact (S8.1), one participant said s/he would stop the robot by pushing the emergency button. Concerning interpersonal power, a significant difference occurred with the robot being rated as more powerful when employing negative compared to neutral and positive strategies $(F(2,106) = 17.72,\ p < .001,\ \eta_p^2 = .24)$. Especially, for a threatening robot, participants reported that the robot controlled the situation and asserted itself. Summarizing, all conflict resolution strategies were more effective than no strategy. Although the robot employing negative strategies was perceived as more powerful, compliance rates for negative strategies were not higher than for positive or negative strategies. Hence, for the public application context, H1 was confirmed and H3 had to be rejected.

4.2.4 Strategy Evaluation: Acceptance, Trust and Fear

In H2 it was expected that negative strategies would be less accepted and less trustworthy than positive and neutral strategies. Acceptance ratings showed that none of the strategies was more accepted than no strategy (S1.1) (see **Figure 5**). Statistical testing revealed a significant difference in acceptance ratings between negative and neutral strategies and between negative and positive strategies $(F(2,120) = 128.3,\ p < .001,\ \eta_\mathrm{p}^2 = .68)$ with negative strategies being less accepted. No difference between neutral and positive strategies occurred. Negative strategies led to less trust than positive and neutral strategies $(F(2,120) = 93.7,\ p < .001,\ \eta_\mathrm{p}^2 = .61)$. No differences occurred between positive and neutral strategies. Negative strategies were rated to evoke more fear than neutral or positive strategies $(F(2,120) = 87.8,\ p < .001,\ \eta_\mathrm{p}^2 = .59)$. No difference for fear ratings occurred between the neutral and positive strategies. Especially, threat (S6.1), annoyance (S4.1) and physical contact (S8.1) had high fear ratings. Descriptively, humor (S12.1) and empathy (S13.1) were the least trustworthy of the positive strategies and empathy (S13.1) had higher fear ratings than the positive or neutral strategies (but less than negative strategies). The evaluation of the context-specific strategies was as follows. Negative public attention (M = 2.6, SD = 1.1) was rated like the negative strategies and positive public attention (M = 5.1, SD = 1.2) was rated equally to the positive strategy, appeal (S9.1). The same results occurred for trust and fear ratings. Summarizing, as expected in H2, negative strategies were less accepted and less trustworthy than positive or neutral strategies.

4.2.4.1 Conflict Resolution Strategy Acceptance Rated by Agent

H4 expected human-like robots to be more accepted to apply conflict resolution strategies than mechanoid robots. The following strategies were more accepted if uttered by the human agent than by any robot: threat (S6.1) $(F(3,60) = 10.90,\ p < .001,\ \eta_p^2 = .31)$, show benefit (S3.1) $(F(3,43) = 4.10,\ p < .05,\ \eta_\mathrm{p}^2 = .19)$, appeal (S9.1) $(F(2,29) = 5.92,\ p < .01,\ \eta_\mathrm{p}^2 = .28)$, apologize (S11.1) $(F(2,51) = 3.81, p < .05, \eta_\mathrm{p}^2 = .15)$, and trigger empathy (S13.1) $(F(2,40) = 5.80,\ p < .01,\ \eta_\mathrm{p}^2 = .23)$. In contrast, the following strategies were more accepted by Roomba compared to all other agents: no strategy (S1.1) $(F(2,51) = 3.45,\ p < .05,\ \eta_\mathrm{p}^2 = .13)$, approach (S7.1) $(F(2,38) = 3.50,\ p < .05,\ \eta_\mathrm{p}^2 = .15)$, and physical contact (S8.1) $(F(2,27) = 5.29,\ p < .05,\ \eta_\mathrm{p}^2 = .26)$. In conclusion, human-like robots were not more accepted to use conflict resolution strategies. As expected in H4, negative strategies were more accepted when applied by a mechanoid robot than by all other robots or the human agent.

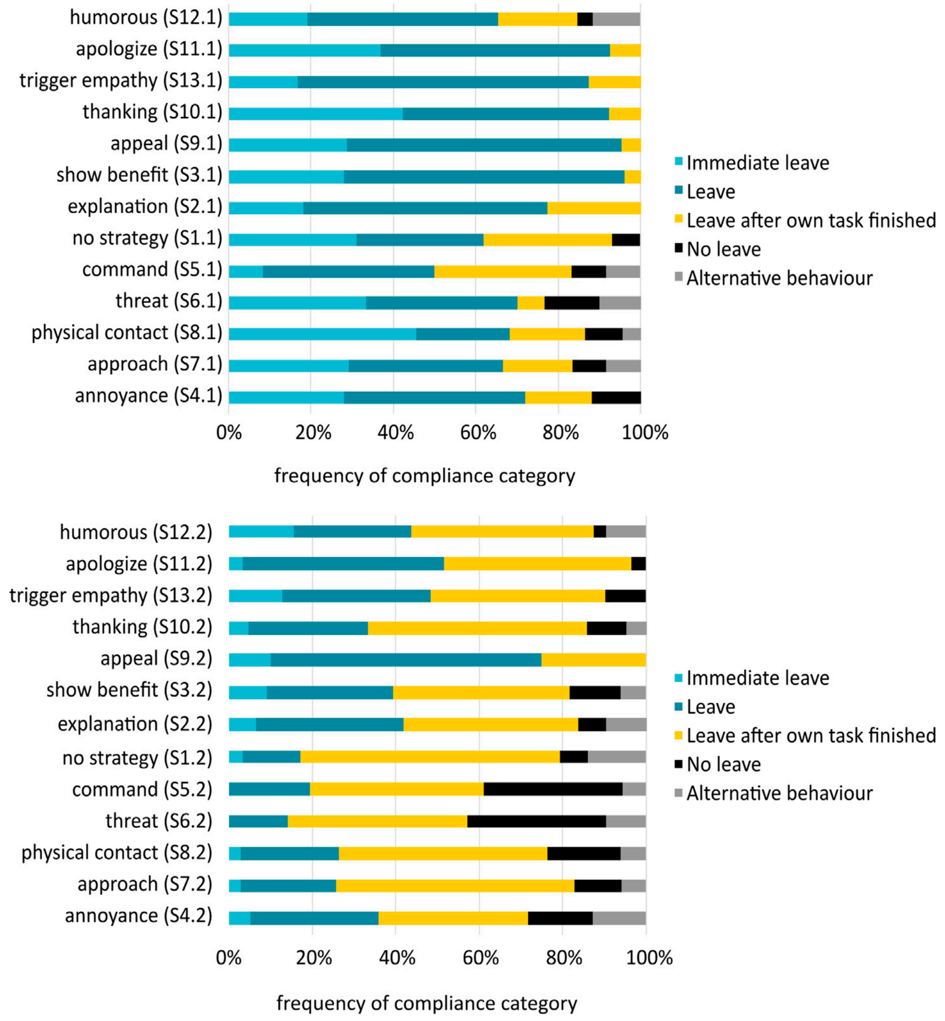


FIGURE 4 | Compliance categories per use context. Public context **(top)** and private context **(bottom)**.

4.2.5 Influences on Strategy Acceptance and Compliance

To explore whether acceptance and compliance are influenced by strategy ratings, correlations were examined. Acceptance correlated highly positively with politeness and trust, as well as moderately negatively with intensity and fear (see **Table 8**). As can be seen in **Table 9**, compliance and interpersonal power were positively correlated but compliance and acceptance did not correlate in the public application context. Strategy intensity and compliance correlated only for the negative strategies. Three stepwise linear regressions with trust, fear of agent behaviour, politeness and interpersonal power as potential predictors on strategy acceptance (negative, neutral, positive) were performed. Politeness and trust transpired as significant predictors for the acceptance of negative, neutral and negative strategies (see **Table 10**). Linear regressions with robot or user characteristics did not produce valuable, predictive models for strategy acceptance. For compliance, an ordinal regression was performed with power, fear, trust and politeness. Compliance with negative strategies could be significantly predicted by interpersonal power (β = 1.39, p < 0.001, CI [0.75; 2.0]) which could explain 36% of compliance variance (Nagelkerke Pseudo R^2 = 0.36). If a participant were to increase his interpersonal power rating by one point, his ordered log-odds of being in a higher compliance category would increase by 1.39 (odds ratio = 4.0). Hence, the higher the perceived interpersonal power was, the more compliant the participants were when the agent applied negative strategies. Positive and neutral strategies showed the same pattern with interpersonal power as significant predictor of compliance but prerequisites were not met. Predictions with robot or user characteristics did not yield valid models. Concluding, the strategy acceptance could be predicted by politeness and trust, indicating that when participants rated the negative strategy as more polite and trustworthy they

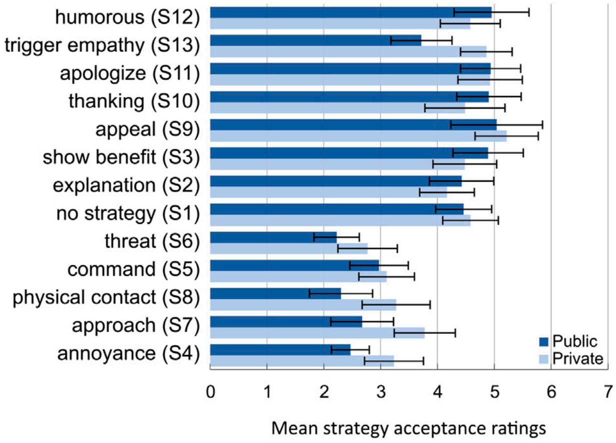


FIGURE 5 | Acceptance ratings per strategy and use context. Error bars indicate ±2 standard errors of the mean.

accepted it more. Participant's compliance with negative strategies was influenced by interpersonal power.

4.2.6 Summary of Results

Concerning compliance, all strategies were more effective in achieving compliance than no strategy (S1.1), except for command (S5.1). Compliance could be predicted by the perceived interpersonal power.

All negative strategies were less accepted than no strategy (S1.1). Cognitive and polite strategies were equally accepted as no strategy (S1.1). Command (S5.1), humor (S12.1) and empathy (S13.1) were neither effective nor accepted. Threat (S6.1) was only accepted for humans but the mechanoid robot Roomba was accepted to use physical strategies (S7.1, S8.1). Evaluative strategy ratings like politeness and trust were significant predictors for strategy acceptance.

5 STUDY 2

5.1 Method

5.1.1 Sample

Forty-eight participants were recruited via email, social media, and flyers on campus. Fifty participants were recruited by a professional online recruiter. Four participants had to be excluded due to video display issues and one due to answer tendencies. The final sample size was $N = 93$. University participants received either course credit or a shopping voucher as compensation. The professionally recruited participants were compensated monetarily.

5.1.2 Study Design

The second online study addressed the private household as an application context for assertive service robots. The study followed a block design where participants saw five out of fifteen conflict resolution strategies. The strategies were implemented in blocks of five negative, three neutral and seven positive strategies. As the context-sensitive strategies (foot-in-the-door (S15.2a) and thanking submissive (S15.2b)) were both positive in valance, an unequal number of negative and positive strategies resulted. The online program randomly assigned two out of five negative, one out of three neutral and two out of seven positive strategies. Not all participants saw all strategies due to test economy. Each strategy was on average rated by 32 participants.

TABLE 8 | Summary of correlations with acceptance.

			Strategy		
		Study	Negative	Neutral	Positive
Acceptance	Trust	Public	0.76	0.71	0.69
		Private	0.77	0.66	0.77
	Fear	Public	−0.47	−0.64	−0.66
		Private	−0.64	−0.47	−0.67
	Politeness	Public	0.76	0.70	0.82
		Private	0.91	0.89	0.83
	Intensity	Public	−0.40	0.46	−0.41
		Private	−0.56	−0.28	−0.28

All correlations were significant on p < .01. Power did not correlate with acceptance.

5.1.3 Human–Robot Goal-Conflict Scenario

The scenario was set in the participant's kitchen where s/he would host a party at home in 15 min. For that, the participant would need to prepare something in the kitchen for the party while it would be important that the robot/person would clean the kitchen before the party started. During preparation, the robot/person would begin to vacuum the kitchen and the participant would be in the way of that process. The participant was then instructed to choose how to behave (see Study 1).

5.1.4 Conflict Resolution Strategies

Applied conflict resolution strategies for both use cases were kept similar (with adapted context-sensitive wording) with four exceptions (see **Table 2**): no strategy (S1.2), foot-in-the-door (S15.2a), thanking submissive (S15.2b) and thanking dominant

TABLE 9 | Summary of correlations with compliance.

			Strategy		
		Study	Negative	Neutral	Positive
Compliance	Acceptance	Public			
		Private	0.39**	0.40**	0.46**
	Trust	Public			
		Private	0.29**	0.40**	0.42**
	Fear	Public			
		Private			−0.26*
	Interpersonal Power	Public	0.61**	0.55**	0.34**
		Private	−0.44**	−0.37**	−0.49**
	Politeness	Public			
		Private	0.32**	0.39**	0.40**
	Intensity	Public	−0.27*		
		Private		−0.35**	

Only significant correlations are shown.
**p < 0.05*
***p < 0.01.*

(S11.2). These strategies were adapted because of lessons-learned from Study 1 or added for a more complete investigation of possible conflict resolution strategies. As adaption to the private context, the baseline strategy (S1.2) included a verbal utterance. The agent uttered the sentence *'I would like to continue to vacuum the kitchen'* and waited. This sentence preceded all other strategies to create transparency regarding the agent's intentions. Another lesson-learned from the participants' comments to the strategies in Study 1, was adapting the wording of the strategy *thanking* (S11.1). In Study 1, the wording of thanking was criticized for being too dominant. Hence, in Study 2 both forms of thanking were compared: submissively (S15.2b) and dominant (S11.2). The *foot-in-the-door technique* (S15.2a) was only applied in the private context. In the public context, this technique did not seem feasible as no small and real request could be formulated to match the private context's (i.e. asking to leave the train station was unsuitable).

5.1.5 Robots and Human Agents

Participants saw videos of three robots: a humanoid service robot TIAGo (PalRobotics), a small vacuum cleaning robot Roomba (iRobot) and a humanoid robot Pepper (SoftBanks) (see **Figure 2**). The robot video's order was randomized. The videos of Roomba and Pepper were the same as in Study 1. Each video lasted between five and 14 s and depicted the robot driving with robot-specific speed towards the viewer in a neutral hallway. Each video was shown twice and participants could not stop or replay the video. After each video, the participant had to confirm the correct video presentation (exclusion criteria). Stimuli videos can be found in the supplementary material along with a screen record of the video presentation in the online survey. Videos and the sketch of each robot were presented as in Study 1. Additionally, the human agent's social role (companion vs. employee) was manipulated to receive a reference value for the robot and strategy ratings based on power asymmetry (companion on equal power level, employee as subordinate). Hence, two human agents were selected: a household member and a domestic help. Both human agents were not introduced with videos to not influence the participants. Instead, the participant was asked to specify which household member s/he imagined during the interaction. The majority of the participants imagined interacting with their partner/spouse (40%) or their flatmate (27%). Summarizing, Study 2 comprised three robots and two human agents.

5.1.6 Study Procedure and Data Analysis

The procedure was identical to Study 1, except for the personality questionnaires. For the private context, where personalizing

TABLE 10 | Regression coefficients for the prediction of strategy acceptance.

Study	Strategy	Predictor	B	Standardized beta	95% CI for B	Model fit R^2 adjusted	R^2 change if Trust is included	Collinearity Tolerance	Statistics VIF
1	Negative	(Intercept)	2.50		[1.83; 3.16]	0.72	0.16	0.70	1.43
		Politeness	0.52	0.49	[0.34; 0.71]				
		Trust	0.39	0.49	[0.25; 0.52]				
	Neutral	(Intercept)	2.10		[0.97; 3.18]	0.57	0.07[a]	0.49	2.04
		Trust	0.47	0.44	[0.21; 0.73]				
		Politeness	0.32	0.38	[0.12; 0.53]				
	Positive	(Intercept)	2.60		[1.64; 3.62]	0.70	0.04	0.55	1.82
		Politeness	0.68	0.64	[0.47; 0.88]				
		Trust	0.32	0.27	[0.09; 0.55]				
2	Negative	(Intercept)	−0.11		[−0.52; 0.30]	0.85	0.03	0.46	2.17
		Politeness	0.73	0.74	[0.61; 0.85]				
		Trust	0.30	0.24	[0.15; 0.46]				
	Neutral	(Intercept)	0.29		[−0.22; 0.79]	0.80	0.02	0.61	1.65
		Politeness	0.74	0.78	[0.63; 0.86]				
		Trust	0.19	0.18	[0.06; 0.31]				
	Positive	(Intercept)	0.12		[−0.42; 0.65]	0.79	0.11	0.60	1.67
		Politeness	0.47	0.57	[0.37; 0.57]				
		Trust	0.49	0.42	[0.34; 0.63]				

[a] *When politeness was included.*

interaction strategies is possible, personality questionnaires regarding general personality traits, conflict type and dispositional empathy were assessed (see **Table 5**). Additionally, the ascribed social role of the robot (e.g. companion, colleague, tool) was assessed as a manipulation check by an open question, followed by a selection of nine potential roles). These additions to the study procedure led to a longer, average study duration of 45 min. Data analysis was similar to Study 1.

5.2 Results

5.2.1 Manipulation Checks

5.2.1.1 Robot Ratings

Participants rated the robots with regard to humanness, uncanniness, power of impact, the potential to produce fear and authority (see **Figure 3**). It was expected that humanoid robots would be perceived more human-like and that larger robots would be perceived as having more power of impact and hence producing more fear. TIAGo was rated as the most uncanny $(F(2, 184) = 75.1,\ p < .001,\ \eta_p^2 = .45)$ and authoritarian robot $(F(2, 184) = 38.5,\ p < .001,\ \eta_p^2 = .30)$. TIAGo and Pepper were rated equally with regard to power and evoked fear. Pepper was rated the most human-like $(F(2, 156) = 32.7,\ p < .001,\ \eta_p^2 = .26)$ whereas Roomba was rated the weakest $(F(2, 169) = 96.9,\ p < .001,\ \eta_p^2 = .51)$ and most mechanical looking robot (see **Figure 3** bottom). For TIAGo and Pepper, the most named social role was employee/butler (22% each). For Roomba, 26% of participants perceived it as having no social role. Twenty-three percent of participants perceived it as a tool and 22% as helper. Summarizing, TIAGo was rated as uncanny, Pepper as the most human-like and Roomba as the most mechanical-looking robot. Both humanoids were perceived as a butler, whereas Roomba was mainly perceived as a tool and as having no social role.

5.2.1.2 Strategy Ratings

To test whether the strategies produced the intended affect and politeness perception, participants rated the strategies concerning valence, intensity and politeness. Strategies that were considered to be negative in valence were rated significantly more negative in valence than the neutral and positive strategies $(F(2, 184) = 46.3, p < .001, \eta_\mathrm{p}^2 = .34)$. Regarding single strategies, more positive strategies than expected were rated as neutral. Approach (S7.2) was not rated as a negative strategy. However, no strategy was rated as very positive (see **Table 7**). Negative strategies were rated as more intense than neutral and positive strategies $(F(2, 157) = 20.7,\ p < .001,\ \eta_\mathrm{p}^2 = .18)$. No difference between positive and neutral occurred. The negative strategies were perceived as more rude than the positive strategies $(F(2, 184) = 48.3,\ p < .001,\ \eta_\mathrm{p}^2 = .34)$. Especially, annoyance (S4.2), command (S5.2), threat (S6.2) and physical contact (S8.2) were rated as the most intense and as the rudest strategies.

5.2.2 Strategy Effectiveness: User Compliance and Interpersonal Power

It was expected that all strategies were more effective than no strategy (H1) and that negative strategies would lead to more compliance than positive and neutral strategies (H3). All strategies were more effective in producing compliance than no strategy (S1.2) (except for threat (S6.2)) (see **Figure 4**), hereby confirming H1. The ANOVA revealed a significant difference in compliance with negative, positive and neutral strategies $(F(2, 164) = 25.0,\ p < .001,\ \eta_\mathrm{p}^2 = .23)$. All post-hoc tests were significant. Concerning the context-specific strategies, the following compliance rates (sum of compliance rates for 'immediate leave' and 'leave') emerged. The foot-in-the-door strategy (S15.2a) was as effective as the average positive strategy with a compliance rate of 46%. Thanking dominant (S.10.2) was as effective as the negative strategies with a compliance rate of 26%. The results of the open answers to the participant's behaviour revealed alternative compliance options: For negative strategies, nine participants stated that they would switch off the robot. For the positive and neutral strategies, four participants indicated that they would tell the robot to drive around them. Regarding interpersonal power, no difference occurred for the ratings between positive, negative and neutral or for the single strategies. Summarizing, as negative strategies were neither rated as more powerful nor were more effective than neutral or negative strategies, H3 had to be rejected.

5.2.3 Strategy Evaluation: Acceptance, Trust and Fear

In H2 it was expected that negative strategies would be less accepted and less trustworthy than positive and neutral strategies. Acceptance ratings showed that none of the strategies was more accepted than no strategy (S1.2) but cognitive and polite strategies were equally accepted (see **Figure 5**). The ANOVA revealed a significant difference of strategy acceptance ratings $(F(2, 184) = 44.5,\ p < .001,\ \eta_\mathrm{p}^2 = .33)$. The post-hoc test showed that negative strategies were less accepted than positive (M = −1.63, p < 0.001) or neutral strategies (M = −1.41, p < 0.001) but no difference between neutral and positive strategies occurred. The evaluation of the two context-specific strategies was as follows. The foot-in-the-door technique (S15.2a) (M = 4.5, SD = 1.6) was as accepted as the neutral strategies. Thanking dominant (S10.2) (M = 3.7, SD = 1.7) was less accepted than thanking submissive (S15.2b) as it was rated like the negative strategies. Concerning trust and fear, negative strategies led to less trust than positive and neutral strategies $(F(2, 165) = 34.4,\ p < .001,\ \eta_\mathrm{p}^2 = .27)$. No differences occurred between positive and neutral strategies but appeal led to the highest trust. Negative strategies were rated to evoke more fear than neutral or positive strategies $(F(2, 184) = 36.3,\ p < .001,\ \eta_\mathrm{p}^2 = .28)$. No difference for fear ratings occurred between the neutral and positive strategies. Especially, annoyance (S4.2) and threat (S6.2) led to the highest fear. Summarizing, as expected negative strategies were less accepted and less trustworthy than positive and neutral strategies which confirms H2 for the private context.

5.2.3.1 Conflict Resolution Strategy Acceptance Rated by Agent

H4 expected human-like robots to be more accepted to apply conflict resolution strategies than mechanoid robots. The household member was the only agent accepted when applying the following conflict resolution strategies: threat (S6.2) ($F(3, 111) = 2.80,\ p < .05,\ \eta_p^2 = .06$), appeal (S9.2) ($F(1, 27) = 8.20, p < .01, \eta_p^2 = .30$), trigger empathy (S13.2) ($F(3, 83) = 3.61,\ p < .05,\ \eta_p^2 = .11$), humor (S12.2) ($F(2, 76) = 11.31,\ p < .001,\ \eta_p^2 = .27$), thanking dominant (S10.2) ($F(2, 63) = 3.71,\ p < .05,\ \eta_p^2 = .13$), and foot-in-the-door (S15.2a) ($F(2, 53) = 4.12,\ p < .05,\ \eta_p^2 = .14$). Only the household member was accepted to express emotional or social conflict resolution strategies. Contrary to expectations in H4, no strategy was more accepted if uttered by a robot regardless of human-likeness. However, most of the strategies were equally accepted for the robots and the domestic help.

5.2.4 Influences on Strategy Acceptance and Compliance

Correlations were examined to explore influences on acceptance and compliance. As can be seen in **Table 8**, acceptance correlated highly positively with politeness and trust, and moderately negatively with intensity and fear. Acceptance and compliance did correlate moderately positively as did politeness and compliance (see **Table 9**). However, compliance and interpersonal power were moderately negatively correlated. Three stepwise linear regressions with trust, fear of agent behaviour, politeness and interpersonal power as potential predictors on strategy acceptance (negative, neutral, positive) were performed. Politeness and trust transpired as significant predictors for the acceptance of negative, neutral and negative strategies (see **Table 10**). Hereby, politeness explained most of the variance of acceptance (see **Table 10**, R^2 changes). Linear regressions with robot or user characteristics did not produce valuable predictive models for strategy acceptance. For compliance, an ordinal regression was performed with power, fear, trust and politeness. Compliance with positive strategies could be significantly negatively predicted by interpersonal power ($\beta = -1.42, p < 0.001$, CI [−1.99; −0.86]) which could explain 44% of compliance variance (Nagelkerke Pseudo R^2 = 0.44). If a participant were to increase his interpersonal power rating by one point, his ordered log-odds of being in a higher compliance category would decrease by 1.42 (odds ratio = 0.24). Hence, the higher the perceived interpersonal power was, the less likely participants' compliance was when the robot applied positive strategies. Negative and neutral strategies showed the same pattern with interpersonal power as significant predictor of compliance but model assumptions were not met. Also predictions with robot or user characteristics on compliance did not yield valid models. Summarizing, acceptance and compliance were positively associated. Higher ratings of strategy intensity and perceived fear resulted in lower acceptance ratings. Strategy acceptance could be predicted by politeness and trust, indicating that when participants rated the negative strategy as more polite and trustworthy they accepted it more. Compliance was positively associated with strategy politeness ratings and negatively with interpersonal power. Hence, if participants rated the strategy as more polite they were more compliant. The more powerful the robot was rated, the less compliant they were.

5.2.5 Summary of Results

All strategies were more effective in achieving compliance than waiting (S1.2), except for command (S5.2) and threat (S6.2). The latter two even led to reactance with about a third of participants not complying. Threat (S6.2) was rated as the least trustworthy and together with annoyance (S4.2) as the two most fearsome strategies. Regarding acceptance, all negative strategies, except for approach (S7.2), were rated as less acceptable than waiting (S1.2) but cognitive (S2.2, S3.2) and polite strategies (S9.2–11.2) were equally accepted. Regarding the agent employing the strategies, no strategy was more accepted if uttered by a robot. Especially, negative strategies (S4.2 - S8.2) and emotional strategies (S12.2, S13.2) were only accepted for the household member. Regarding influences on acceptance and compliance, acceptance was connected to politeness, trust, and fear. Compliance was negatively associated with interpersonal power and politeness in the private context. Compliance and acceptance correlated moderately.

6 DISCUSSION

The aim of this study was to develop and test conflict resolution strategies for service robots to achieve compliance with a robot's request in an accepted way. For this, psychological principles were transferred to HRI to develop conflict resolution strategies. The strategies were systematically tested in two online studies in two application contexts for service robots: public and private space. Hereby, the strategy classification into three valence categories allowed for systematically testing as each participant rated the same amount of negative, neutral and positive strategies. The results showed that neutral and positive conflict resolution strategies were accepted and effective in achieving compliance with a robot's request. Negative strategies were more controversial as user acceptance and compliance were dependent on robot type and application context. Negative strategies like command (S5.2) and threat (S6.2) even led to user reactance. For the public context, influences on strategy acceptance and compliance could be found. Whereas acceptance was predicted by politeness and trust, compliance was predicted by interpersonal power.

Based on the results, two hypothesis could be accepted and one had to be rejected. Regarding the conflict resolution strategies, it was expected that they would be more effective than no strategy (H1). This was true for both application contexts (except for command and threat). Hence, H1 was supported. However, not all strategies can be recommended to be pursued further, as will be described below. Regarding negative strategies, it was assumed on the basis of the human-power asymmetry that strategies with high interpersonal power of the robot would be evaluated negatively in terms of acceptance and trust (H2), but would lead to more compliance (H3). For both application contexts, negative strategies like commanding (S5) were found to be less accepted and less effective in achieving compliance than positive strategies. Hence, H2 (acceptance, trust) was supported and H3 (compliance) had to be declined. Negative strategies even led to psychological reactance with about one-tenth to one-third of participants in both application contexts indicating that they intentionally disobeyed. Reactance was more common in the private than in the public application context. Only here, a positive correlation between politeness and compliance occurred, indicating that the more rude a request was perceived the less likely compliance was. This was mirrored in the correlations between interpersonal power and compliance. Whereas compliance and interpersonal power were highly correlated in both application contexts, only in the private context, the correlation was negative. Hence, the user did not comply even if s/he rated the robot as more powerful than him/herself. This illustrates, as expected, the higher effect of the power asymmetry in the private context. The reactance found in this study has been found in previous work (Roubroeks et al., 2010; Ghazali et al., 2018). Only in the private context, compliance and acceptance ratings were moderately, positively correlated. This might hint to the possibility that strategy acceptance might be more important in the private application context than in public. In the private context, where one has robot control and authorization, acceptance guides the compliance decision. In the public context, one might comply although not accepting the robot's request because one feels in a weaker position and publicly observed.

In H4 it was expected that human-like robots would be more accepted to apply positive and neutral conflict resolution strategies compared to mechanoid robots. In both application contexts, it was more accepted if the human uttered the negative strategy threat (S6), the positive strategy appeal (S9) or the human-specific strategy empathy (S13) than if a robot did. As expected, the mechanoid robot Roomba was more accepted to use negative conflict resolution strategies than Pepper in public. In the private context, no strategy was more accepted if uttered by a robot regardless of human-likeness. Hence, H4 was only partially confirmed. However, most of the strategies were equally accepted for the robots and the domestic help. Only the household member with the assumed same social status as the participant was accepted to express emotional or social conflict resolution strategies. This may indicate a greater influence of social status on the acceptance of certain conflict resolution strategies in the private context than the human-likeness of the robot. For all other strategies in both contexts, no difference in acceptance occurred between robots and humans which shows the potential of robotic conflict resolution strategies. Hereby, more research is needed to determine the appropriate set of conflict resolution strategies per robot type and application context.

Apart from the hypotheses, a research question was formulated that concerned the differences between application contexts regarding strategy acceptance and effectiveness. Indeed differences between the contexts showed. For the private context, all positive strategies were rated as more polite than no strategy (S1) which was the opposite in the public context. Additionally, all negative strategies, except for command (S5.2), were more accepted in the private application context. Although negative strategies were less accepted in the public context, compliance rates for negative strategies were higher compared to the private context. Interestingly, human-robot power asymmetry influenced the prominent way of compliance. Whereas in public (assumed human-robot power equality), participants' prevalent reaction was to comply (not immediately), they favored finishing their task first in the private context (assumed owner superiority). In a study which tried to elicit helping behaviour from participants who were occupied with a secondary task showed that people preferred to help after they had finished their task instead of interrupting it (Fischer et al., 2014).

Differences between application contexts also appeared for effective strategy mechanisms. Hereby, cognitive and polite strategies were most accepted and successful findings regarding social strategies were mixed. Authority-based strategies (i.e. S5 command and S6 threat) were neither accepted nor effective. This was also true for strategies using negative reinforcement (S4 annoyance) and negative social influence (S14.1b negative public attention). In contrast, positive social strategies using a sequential-compliance technique (S15.2a foot-in-the-door) or positive social influence (S14.1a positive public attention) were accepted and effective. Therefore, if an assertive robot makes use of social influence, it should be in a positive manner to avoid negative effects of human-robot power asymmetry. Concerning emotional strategies, empathy (S13.1), but not humor (S12.1), was less accepted in the public context. Empathy (S13.1) was rated as less trustworthy and more fearsome than other positive or neutral strategies in the public context. As the robot in the public context might be perceived as equal due to its social role, trying to elicit empathy for its situation (i.e. appearing weaker) could contradict the role assumption. Just as it is considered inappropriate for a cleaner to address a passer-by on a personal level, the same could apply to an autonomous service robot. Similarly, in the private context, emotional strategies (S12.2, S13.2) were only accepted for the household member but not for any robot. Regarding physical strategies, they were more accepted in the private than in the public context. As physical strategies emphasize the robot's embodiment, they are likely connected to fear of the robot. Indeed, in the public context, physical strategies (S7.2, S8.2) were rated as more fearful than in the private context. A higher fear in the public context might be explained by a lack of prior information about the robot's function and capabilities compared to the public. This is also mirrored in the interaction between strategy mechanism and robot type in public. Both physical strategies (S7.1, S8.1) were more successful for a small, non-threatening robot (Roomba) compared to other robots and the human agent. Naturally, if the users do not fear that an assertive robot might harm them, the robot is more accepted. This is in line with previous studies regarding robot size and perceived power of impact (Young et al., 2009; Jost et al., 2019). Hereby, pre-information and transparency will be important in the future to ensure that an assertive robot, regardless of size and strength, will never use force. In the private context, a robot respecting the user's personal space (S7.2 approach) was more accepted than a close approach (5 cm in the presented study as in S8.2 physical contact). As in previous findings a positive effect on compliance was found with a minimum distance of 0.6 m (Mutlu, 2011; Chidambaram et al., 2012), our implementation was probably too close for comfort. Since the presented study was conducted online, the results regarding the physical mechanisms for robot conflict resolution strategies require further confirmation. Summarizing, application context differences regarding effective mechanisms suggest that robotic conflict resolution strategies need to be applied context-sensitively to be useful.

Having established strategies' acceptability and effectiveness, a first test of influencing factors on those variables was performed. In both application contexts, acceptance ratings could be predicted by politeness and trust ratings. Similar to human negotiations (Pfafman, 2017), perceived politeness and trust were influential on strategy acceptance in both contexts. This might explain why integrative robot conflict resolution strategies were more effective and accepted than distributive strategies. Similarly, in human negotiations integrative strategies are preferred if trust between negotiators is high (Kong et al., 2014). Therefore, integrative strategies seem more promising in HRI than distributive conflict resolution strategies for both application contexts. For both application contexts, interpersonal power could predict compliance but the influence differed. In the public context, compliance with negative strategies could be positively predicted by the higher interpersonal power of the robot. Naturally, higher robot power led to higher compliance. In contrast, in the private context, compliance with positive strategies was negatively predicted by higher interpersonal power. Hence, although the robot was rated as more powerful, the participants were still less likely to comply. Once, more this could represent the higher impact of the power asymmetry in the home context. Here, even positive strategies might be perceived as inappropriate. This is also supported by the finding that no robotic conflict resolution strategy was highly accepted (average of five on a 7-Point Likert Scale). Therefore, in the home context, the robot user's personal assessment of the human-robot power asymmetry is an important factor that needs to be considered for real-world applications. User variables regarding general personality, conflict type, dispositional empathy, demographics, robot experience/ownership or negative attitudes towards robots could not predict strategy acceptance or compliance. Potentially, a correlative design with a larger sample size has more potential to determine if user characteristics influence human-robot goal conflict resolution as they do in human-human interactions. Summarizing, differences were found between the developed conflict resolution strategies regarding compliance, acceptance and trust between the use contexts and were influenced by perceived interpersonal power and politeness. In addition to previous studies (Saunderson and Nejat, 2019), the presented findings can now serve as a basis for the application and further development of robotic conflict resolution strategies. Recommendations for the public and private application context are presented below.

6.1 Practical Implications

Concerning a real-world application of robot assertiveness, conflict resolution strategies could have the potential to render service robots in public and private more useful if such robot behaviour is accepted. Based on the theoretical background and empirical findings, we would like to present the following recommendations regarding acceptable and effective conflict resolution strategies for autonomous service robots.

Recommended conflict resolution strategies for the public application context are:

- Goal explanation (S2.1), showing the benefit of cooperation (S3.1), humor (S12.1), positive public attention (S14.1a), approach (S7.1) (if applied by small robot).

Not recommended for the public context:

- Annoyance (S4.1), command (S5.1), threat (S6.1), physical contact (S8.1), eliciting empathy (S13.1), negative public attention (S14.1b).

Recommended conflict resolution strategies for the private application context are:

- Goal explanation (S2.2), showing the benefit of cooperation (S3.2), approach (S7.2), foot-in-the-door (S15.2a).

Not recommended for the private context:

- annoyance (S4.2), command (S5.2), threat (S6.2), physical contact (S8.2).

Polite strategies like appeal (S9), thanking (S10) and apologizing (S11) can be used in addition to the conflict resolution strategies. Future studies could examine if a combination of assertive strategies with polite strategies is more accepted and effective than a single strategy approach. As in human negotiations, politeness could reduce the face threats posed by assertive strategies and make them more acceptable (Pfafman, 2017). Hereby, learning from psychology, an escalating manner might be feasible: applying assertive strategies after polite, cooperative strategies have failed might be more acceptable (Preuss and van der Wijst, 2017). For this, combining cognitive mechanisms like goal explanation (S2) and showing benefit (S3) with polite strategies (S9–S11) could be especially beneficial as both were effective and accepted in both application contexts. In practice, one possible implementation of conflict resolution strategies for the private context could be: first appeal (S9.2), then show the benefits of cooperation (S3.2) and finally, if the participant has not complied, try the foot-in-the-door technique (S15.2a). Future studies can then test if strategy combinations are more effective and acceptable than single strategy approaches. Hereby, observed application context and robot differences regarding strategy effectiveness and acceptability require a context-sensitive and robot-specific strategy development. Whereas cognitive and polite strategies seem feasible for both contexts, emotional and physical strategies were more acceptable for the private context. However, if a small mechanoid robot applies physical strategies (S7.1, S8.1), they could also be accepted in public. Regarding compliance, a robot using high power strategies (e.g. S5 command and S6 threat) can lead to reactance, especially in the private application context. In general, compliance with a robot's request should be expected to be lower in the private application context than in public due to power asymmetry. Hereby, for real-world applications of assertive service robots at home it might be important to assess the user's preferences regarding the robot's autonomy and assertiveness level. For instance, if the service robot is delivered, the user could answer the respective questions and the robot's level of robot assertiveness is personalized accordingly. Although some might deny robot assertiveness at the first assessment, it is conceivable that they will be convinced by time as conflict situations occur where the robot will be ineffective if it always defers to the user. Hereby, also trust and politeness will decide about the long-term acceptance of robot assertiveness. For the public context where personalizing is not feasible robot assertiveness should only be applied purposefully and in moderation to solve human-robot goal conflicts. This includes that before issuing the request in a crowded place, the robot checks whether the person addressed actually has the possibility to comply with the request (e.g. space and time for evasion; disability) in order not to disturb passers-by. Situational adaption of robot assertiveness might be key for long-term acceptance of assertive service robots in public. Finally, the ethical implications of robot assertiveness similar to persuasive robots (Chidambaram et al., 2012) need to be considered. Robot assertiveness could be an acceptable and effective form of robot goal achievement as long as it supports goals deemed appropriate by the user and society and never uses violence.

6.2 Strengths and Limitations

This study is the first to develop robot conflict resolution strategies that are based on psychological mechanisms of goal conflict resolution. The theoretical foundation had the advantage of developing a variety of potentially effective strategies which have not been focused in HRI yet and subsequently extends the design scope of robotic interaction strategies. Additionally, systematically considering the psychological mechanisms of conflict resolution strategies allowed for a deeper understanding of the results. The combination of two robot application contexts and different robot types (large, small, humanoid, mechanoid) allowed more precise statements to be made about the specific effectiveness of the strategies and their acceptance. This way, the study was able to investigate the specific effects conflict resolution strategy combinations with different robot types and application contexts. The online study format allowed for a text-based strategy presentation without the influence of the real-world implementation into a certain robot prototype (e.g. appearance, specifications, speech synthesis limitations). This meant that the strategy effect could be investigated without biases added by the implementation. When setting up the online studies, standardization of study material was emphasized, by amongst others, ensuring that the robot videos were of the same length, assessing whether the participants got the video displayed correctly, and using validated questionnaires where possible. Manipulation checks regarding robot ratings were successful.

Although the presented studies have provided insights into the acceptance and effectiveness of robot assertiveness, some limitations of the study have to be considered. The extensive testing of fifteen conflict resolution strategies per application context meant that not all participants saw all strategies. This limited the statistical power but, at the same, time diminished potential respondent fatigue. Regarding internal validity, standardization of strategies was difficult with regard to sentence length. Polite speech is naturally more indirect and lengthy as it tends to paraphrase and embellish (Danescu-Niculescu-Mizil et al., 2013). Strategy phrasing has shown to be essential regarding this study's findings. Thanking dominant (S10.2) was perceived as a negative strategy compared to thanking submissive (S15.2b) which was positively evaluated. Hence, it was reasonable to differentiate between thanking dominant and submissive in Study 2. Consequently, the phrasing for a thankful strategy has to be chosen carefully (present tense vs. subjunctive). For the comparison between the application contexts, it has to be noted that the presented results can only provide first evidence regarding context differences. As the application context was not implemented as an independent variable and the robots differed, further studies are needed which compare both application contexts directly. Although the strategy classification into three valence categories allowed for systematically testing participants' ratings differed from the expected affective evaluation. Some of the positive strategies were rated more neutrally than expected and none was rated very positively. The categorization based on the human-power asymmetry should not be seen as final but as a working hypothesis that allows for systematically testing. However, it shows the relevance of assessing participant's perception of strategy valence for future testing of robotic conflict resolution strategies. Finally, as the evaluation was conducted online, external validity might be limited. As only the intention to comply could be measured and videos cannot replace real world encounters, lab and field experiments are needed to replicate results. This holds especially true for physical strategies which might have been difficult to imagine although they were described in relativity to the participants position (e.g. until the robot touches your luggage). Limitations regarding immersion seem likely but that the robot behaviour could trigger reactance and that some strategies (e.g. threat and command) were not even accepted in an online setting with imagined interaction indicates the psychological reality of the participants during the study. It has also been shown in previous HRI studies that imagined interaction with a robot does resemble real HRI with regard to acceptance of the robot, participant's behaviour toward the robot (Wullenkord and Eyssel, 2019) and negative attitudes towards the robot (Wullenkord et al., 2016).

Therefore, guided imagined interactions seemed to be reasonable for conducting preliminary evaluations of the developed strategies. The intention behind the online format was not to replace real-world testing but to detect strategies that might already be rejected in an imagined situation (which was indeed the case for threat and command) and eliminate them from future research agendas regarding acceptable and effective robotic conflict resolution strategies. Then, for real-world testing, it can be focused on the final best-accepted strategies. Beyond the limitations of online testing, the external validity of the results is questionable as the conflict resolution strategies were examined in a specific situation with specific robots. Therefore, future work might aim to clarify the extent to which results can be generalized to different situations, robots and contexts.

6.3 Future Work

Future studies are needed to determine factors that render some robotic conflict resolution strategies more acceptable and effective than others. Hereby, robot, human and situational influences need to be considered. On the robot side, the strategy implementation must be skilfully implemented in terms of speech (e.g. tone of voice), gestures and proximity. Appropriate expression of assertiveness in human conflict resolutions is considered a communication skill that is not trivial to acquire (Pfafman, 2017). For this, it seems reasonable to rely on psychological research not only for strategy development but also for implementation, e.g. training programs to promote appropriate assertiveness at work (Thacker and Yost, 2002; Wilson et al., 2003; Nakamura et al., 2017). Additionally, future work is needed to determine appropriate conflict resolution strategies for more robot types (e.g. androids) and sizes (e.g. miniature, man-sized) which were not represented in the presented studies. Potentially, with an even more varied set of robots than used in the presented studies, robot characteristics like humanness, power of impact and authority might turn out as moderators for strategy effectiveness and acceptance.

On the human side, user personality, robot attitudes and pre-experience, as well as culture, are likely to be of importance for strategy acceptance and effectiveness as they are influential in human negotiations. Here, general personality traits (BIG5, Costa and McCrae, 1985) and specific conflict-related traits such as the conflict type (ROCI-II, Rahim, 1983) have shown to determine individual conflict behaviour (Rahim, 1983; Park and Antonioni, 2007). An integrating style was positively associated with Agreeableness and Extraversion (Park and Antonioni, 2007). Dominating personalities use distributive conflict resolution strategies (Rahim, 1983) and are positively associated with Extraversion but negatively with Agreeableness (Park and Antonioni, 2007). Conceivably, the robot's strategy has to match the user conflict personality to be effective and accepted. If a dominating negotiator is confronted with an assertive robot, the robot might be less acceptable than if the robot had applied the strategy to a person with an obliging conflict style. In addition, negative attitudes and fears about robots could negatively influence the acceptance of and compliance with assertive robots, since such individuals already tend not to accept non-assertive robots (De Graaf and Allouch, 2013b; Ghazali et al., 2020). Negative attitudes and state anxiety have also shown to negatively influence trust in HRI (Miller et al., 2020). Culture is an additional influence that needs examination in future work. Cultural expectations shape expectations regarding politeness and assertiveness (Lee et al., 2017). Assertiveness must be considered appropriate (e.g. to context and culture), otherwise it can be perceived as aggressive (Pfafman, 2017). An assertive robot might be acceptable in Eurasian countries but could be considered as inappropriate and rude in Asian countries. For Germans and Chinese this has been shown for assertive communication strategies of a small autonomous delivery robot towards pedestrians (Lanzer et al., 2020). Consequently, the presented findings need further confirmation in different samples. Summarizing, future studies are needed to determine the influences of user characteristics on the acceptance of robot assertiveness. Findings could then be used to personalize the robot in the home setting as it has been suggested with other robot characteristics (Ligthart and Truong, 2015).

Situational influences on strategy acceptance and effectiveness are likely to be the conflict scenario (e.g. emergency situations), other application contexts (security robots), repetition and habituation. Apart from the presented scenarios, robot assertiveness could be especially useful for emergency situations. In the public context, for example, security robots might help during an evacuation and might need to be assertive to gain people's trust and compliance in such a stressful, chaotic situation. In the private context, a service robot might need to be assertive and call an ambulance in case of a medical emergency. To avoid that the results are possibly distorted by the novelty effect of an assertive robot, it is necessary to test whether repeated interaction changes the participants' attitude and behaviour towards the robot's assertiveness (e.g. habituation, trust building). If the user benefited from the autonomy and effectiveness of the robot in the past and trust was built up through reliable functioning and appropriate robot actions, the acceptance of the robot's assertiveness could increase (Ghazali et al., 2020; Kraus et al., 2019; Kraus et al., 2020). Similarly, human-robot power asymmetry might be reduced by habituation when assertive robots become an effective and accepted part of our society. This paper represents the first step towards this goal.

7 CONCLUSION

With future dissemination of service robots in public and private spaces, human-robot goal conflicts will arise. To negotiate acceptable outcomes and for efficient task execution, it might be feasible to apply an assertive robot behaviour under certain circumstances. This study explored different conflict resolution strategies, ranging from polite to assertive, to achieve user compliance and acceptance simultaneously in two application contexts, public and private space. The potential of applying robotic conflict resolution strategies to increase intended compliance with a robot's request in an acceptable way was shown. Positive and neutral conflict resolution strategies were acceptable and effective in achieving compliance with a robot's request and should be explored further. Combining strategies based on cognitive mechanisms with politeness seems especially feasible for both application contexts. Only command (S5) and threat (S6) do not seem feasible to be examined further as they were neither effective nor accepted. The perceived interpersonal power of the robot influenced the participants' decision to comply. Trust and politeness were predictive of strategy acceptance. Concluding, if applied context-sensitively and robot-specifically, robotic conflict resolution strategies as an appropriate expression of robot assertiveness have the potential to solve human-robot goal-conflicts effectively and acceptably. This study represents a first step to designing conflict resolution strategies for future assertive robots. Future work is needed to determine factors that render robot assertiveness acceptable for various users, robots and situations.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

All authors were involved in the process. FB lead the ideation, conception, study design, data analysis and writing, JK was substantial to study design and data collection, editing and additional writing, while MB contributed to study design, editing and supervision.

FUNDING

This research has been conducted within the interdisciplinary research project 'RobotKoop' which is funded by the German Ministry of Education and Research (Grant Number 16SV7967).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2020.591448/full#supplementary-material.

REFERENCES

Adam, H., and Shirako, A. (2013). Not all anger is created equal: the impact of the expresser's culture on the social effects of anger in negotiations. *J. Appl. Psychol.* 98, 785–798. doi:10.1037/a0032387

Albert, S., and Dabbs, J. M. (1970). Physical distance and persuasion. *J. Pers. Soc. Psychol.* 15, 265–270. doi:10.1037/h0029430

Argyle, M., and Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, 29, 289–304.

Babel, F., Kraus, J., Miller, L., Kraus, M., Wagner, N., Minker, W., et al. (2021). Small talk with a robot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. *Int. J. Soc. Robot.* doi:10.1007/s12369-020-00730-0

Bartneck, C., and Hu, J. (2008). Exploring the abuse of robots. *Interact. Stud. Soc. Behav. Commun. Biol. Artif. Syst.* 9, 415–433. doi:10.1075/is.9.3.04bar

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* 1, 71–81. doi:10.1007/s12369-008-0001-3

Bechade, L., Duplessis, G. D., and Devillers, L. (2016). "*Empirical study of humor support in social human-robot interaction*", in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 9749. (New York, NY: Springer), 305–316. doi:10.1007/978-3-319-39862-428

Berridge, K. C. (2001). Reward learning: reinforcement, incentives, and expectations. *Psychol. Learn. Motiv.* 40, 223–278. doi:10.1016/S0079-7421(00)80022-5

Betancourt, H. (2004). Attribution-emotion processes in White's realistic empathy approach to conflict and negotiation. *Peace Conflict* 10, 369–380. doi:10.1207/s15327949pac10047

Boardman, A. E., Greenberg, D. H., Vining, A. R., and Weimer, D. L. (2017). *Cost-benefit analysis: concepts and practice*. (Cambridge: Cambridge University Press). doi:10.1080/13876988.2016.1190083

Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatr.* 25, 49–59.

Brett, J., and Thompson, L. (2016). Negotiation. *Organ. Behav. Human Decis. Process* 136, 68–79. doi:10.1016/J.OBHDP.2016.06.003

Brown, P., Levinson, S. C., and Levinson, S. C. (1987). *Politeness: Some universals in language usage*, vol. 4. (Cambridge, UK: Cambridge University Press).

Cann, A., Calhoun, L. G., and Banks, J. S. (1997). On the role of humor appreciation in interpersonal attraction: it's no joking matter. *Humor Int. J. Humor Res.* 10, 77–90.

Castro-González, Á., Castillo, J. C., Alonso-Martín, F., Olortegui-Ortega, O. V., González-Pacheco, V., Malfaz, M., et al. (2016). "*The effects of an impolite vs. A polite robot playing rock-paper-scissors*", in *Lecture Notes in Computer Science (including subseries Lecture Notes in artificial Intelligence and Lecture Notes in Bioinformatics) 9979 LNAI*, 306–316. doi:10.1007/978-3-319-47437-330

Chaiken, S. L., Gruenfeld, D. H., and Judd, C. M. (2000). *Persuasion in negotiations and conflict situations* (San Francisco, CA: Jossey-Bass).

Chidambaram, V., Chiang, Y.-H., and Mutlu, B. (2012). "Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues", in Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, 293–300.

Christenson, A. M., Buchanan, J. A., Houlihan, D., and Wanzek, M. (2011). Command use and compliance in staff communication with elderly residents of long-term care facilities. *Behav. Ther.* 42, 47–58. doi:10.1016/j.beth.2010.07.001

Cialdini, R. B. (2009). *Influence: Science and practice*, vol. 4. (Boston, MA: Pearson education).

Cialdini, R. B., and Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annu. Rev. Psychol.* 55, 591–621. doi:10.1146/annurev.psych.55.090902.142015

Cialdini, R. B., Wosinska, W., Barrett, D. W., Butner, J., and Gornik-Durose, M. (1999). Compliance with a request in two cultures: the differential influence of social proof and commitment/consistency on collectivists and individualists. *Pers. Soc. Psychol. Bull.* 25, 1242–1253. doi:10.1177/0146167299258006

Cohen, T. R. (2010). Moral emotions and unethical bargaining: the differential effects of empathy and perspective taking in deterring deceitful negotiation. *J. Bus. Ethics*. 94, 569–579. doi:10.1007/s10551-009-0338-z

Cormier, D., Newman, G., Nakane, M., Young, J. E., and Durocher, S. (2013). "Would you do as a robot commands? an obedience study for human-robot interaction", in International Conference on Human-agent Interaction.

Costa, P. T., and McCrae, R. R. (1985). *NEO five factor inventory 1989.*

Da-peng, L., and Jing-hong, W. (2017). "Business negotiation skills based on politeness principle", in Asia International Symposium on language Literature and Translation, 232.

Danescu-Niculescu-Mizil, C., Sudhof, M., Dan, J., Leskovec, J., and Potts, C. (2013). "A computational approach to politeness with application to social factors", in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics 1, 250–259. doi:10.1.1.294.4838

Darling, K., Nandy, P., and Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. *Proc. IEEE Int. Work. Robot Hum. Interact. Commun.* 73, 770–775. doi:10.1109/ROMAN.2015.7333675

De Dreu, C. K. W. (2010). "*Social conflict: the emergence and consequences of struggle and negotiation*", in *Handbook of social psychology*. (Amsterdam, Netherlands: University of Amsterdam). doi:10.1002/9780470561119.socpsy002027

De Graaf, M. M., and Allouch, S. B. (2013a). Exploring influencing variables for the acceptance of social robots. *Robot. Autonom. Syst.* 61, 1476–1486. doi:10.1016/j.robot.2013.07.007

De Graaf, M. M., and Allouch, S. B. (2013b). The relation between people's attitude and anxiety towards robots in human-robot interaction. *Proc. IEEE Int. Work. Robot Hum. Interact. Commun*. 628, 632–637. doi:10.1109/ROMAN.2013.6628419

Dillard, J. P. (1991). The current status of research on sequential-request compliance techniques. *Pers. Soc. Psychol. Bull.* 17, 283–288. doi:10.1177/0146167291173008

Drury, J. L., Scholtz, J., and Yanco, H. A. (2003). "Awareness in human-robot interactions", in SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483) (IEEE*)*. 1, 912–918.

Fasola, J., and Matarić, M. J. (2009). "Robot motivator: improving user performance on a physical/mental task", in 2009 4th ACM/IEEE International Conference on Human-robot Interaction (HRI) (IEEE), 295–296.

Fischer, K., Soto, B., Pantofaru, C., and Takayama, L. (2014). "Initiating interactions in order to get help: effects of social framing on people's responses to robots' requests for assistance", in The 23rd IEEE International Symposium on robot and Human Interactive Communication. IEEE, 999–1005.

Fogg, B. J. (2002). *Persuasive technology: using computers to change what we think and do (interactive technologies)*. (Burlington, MA: Morgan Kaufmann). doi:10.1145/764008.763957

Freedman, J. L., and Fraser, S. C. (1966). Compliance without pressure: the foot-in-the-door technique. *J. Pers. Soc. Psychol.* 4, 195.

Gerpott, F. H., Balliet, D., Columbus, S., Molho, C., and de Vries, R. E. (2018). How do people think about interdependence? A multidimensional model of subjective outcome interdependence. *J. Pers. Soc. Psychol.* 115, 716–742. doi:10.1037/pspp0000166

Ghazali, A. S., Ham, J., Barakova, E., and Markopoulos, P. (2020). Persuasive robots acceptance model (pram): roles of social responses within the acceptance model of persuasive robots. *Int. J. Soc. Robot*. 12, 1075–1092. doi:10.1007/s12369-019-00611-1

Ghazali, A. S., Ham, J., Barakova, E., and Markopoulos, P. (2018). The influence of social cues in persuasive social robots on psychological reactance and compliance. *Comput. Hum. Behav*. 87, 58–65. doi:10.1016/j.chb.2018.05.016

Ghazizadeh, M., Lee, J. D., and Boyle, L. N. (2012). Extending the technology acceptance model to assess automation. *Cognit. Technol. Work*. 14, 39–49. doi:10.1007/s10111-011-0194-3

Gilbert, P., and Allan, S. (1994). Assertiveness, submissive behaviour and social comparison. *Br. J. Clin. Psychol.* 33, 295–306.

Gilet, A., Mella, N., Studer, J., and Grühn, D. (2013). Assessing dispositional empathy in adults: a French validation of the interpersonal reactivity index (IRI). *Can. J. Behav. Sci.* 45, 42–48. doi:10.1037/a0030425

Glick, P., DeMorest, J. A., and Hotze, C. A. (1988). Keeping your distance: group membership, personal space, and requests for small favors 1. *J. Appl. Soc. Psychol*. 18, 315–330.

Goetz, J., Kiesler, S., and Powers, A. (2003). "Matching robot appearance and behaviour to tasks to improve human-robot cooperation", in RO-MAN 2003: the 12th IEEE international workshop on robot and human interactive communication, 55–60.

Goldstein, A. P., and Michaels, G. Y. (1985). *Empathy: development, training, and consequences*. (Mahwah, NJ: Lawrence Erlbaum).

Groom, V., and Nass, C. (2007). Can robots be teammates? Benchmarks in human–robot teams. *Interact. Stud*. 8, 483–500. doi:10.1075/is.8.3.10gro

Guadagno, R. E. (2014). "Compliance: a classic and contemporary review," in *Oxford Handbook of social influence*. Editors S. Harkins, W. Kipling, and J. Burger (Oxford: Oxford University Press), 107–127. doi:10.1093/oxfordhb/9780199859870.013.4

Hall, E. T. (1974). *Handbook for proxemic research*. (Washington: Society for the Anthropology of Visual Communication).

Ham, J., and Midden, C. J. H. (2014). A persuasive robot to stimulate energy conservation: the influence of positive and negative social feedback and task similarity on energy-consumption behavior. *Int. J. Soc. Robot*. 6, 163–171. doi:10.1007/s12369-013-0205-z

Hassenzahl, M., Burmester, M., and Koller, F. (2003). "*Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität*", in *Mensch & Computer 2003*. (New York, NY: Springer), 187–196.

Ho, C. C., and MacDorman, K. F. (2017). Measuring the uncanny valley effect: refinements to indices for perceived humanness, attractiveness, and eeriness. *Int. J. Soc. Robot*. 9, 129–139. doi:10.1007/s12369-016-0380-9

Hüffmeier, J., Freund, P. A., Zerres, A., Backhaus, K., and Hertel, G. (2014). Being Tough or being Nice? A meta-analysis on the impact of hard- and softline strategies in distributive negotiations. *J. Manag*. 40, 866–892. doi:10.1177/0149206311423788

Inbar, O., and Meyer, J. (2015). "*Manners matter: trust in robotic peacekeepers*", in *Proceedings of the human factors and Ergonomics society*, 2015, 185–189. doi:10.1177/1541931215591038

Janssen, J. B., van der Wal, C. C., Neerincx, M. A., and Looije, R. (2011). "Motivating children to learn arithmetic with an adaptive robot game", in International conference on social robotics (New York, NY: Springer), 153–162.

Jarrassé, N., Sanguineti, V., and Burdet, E. (2014). Slaves no longer: review on role assignment for human-robot joint motor action. *Adapt. Behav*. 22, 70–82. doi:10.1177/1059712313481044

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cognit. Ergon*. 4, 53–71.

Jost, J., Kirks, T., Chapman, S., and Rinkenauer, G. (2019). "*Examining the effects of height, velocity and emotional representation of a social transport robot and human factors in human-robot collaboration*", in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 11747 (New York, NY: Springer), 517–526. doi:10.1007/978-3-030-29384-0̕31

Kamei, K., Shinozawa, K., Ikeda, T., Utsumi, A., Miyashita, T., and Hagita, N. (2010). "Recommendation from robots in a real-world retail shop", in International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction, 1–8.

Kirst, L. K. (2011). *Investigating the relationship between assertiveness and personality characteristics*. B.S. Thesis.

Kobberholm, K. W., Carstens, K. S., Bøg, L. W., Santos, M. H., Ramskov, S., Mohamed, S. A., et al. (2020). "The influence of incremental information presentation on the persuasiveness of a robot", in HRI 2020: companion of the 2020 ACM/IEEE international conference on human-robot interaction, 302–304.

Kong, D. T., Dirks, K. T., and Ferrin, D. L. (2014). Interpersonal trust within negotiations: meta-analytic evidence, critical contingencies, and directions for future research. *Acad. Manag. J*. 57, 1235–1255. doi:10.5465/amj.2012.0461

Kraus, J., Scholz, D., Stiegemeier, D., and Baumann, M. (2019). The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Hum. Factors*, 62(5), 718–736. doi:10.1177/0018720819853686

Kraus, J., Scholz, D., Messner, E.-M., Messner, M., and Baumann, M. (2020). Scared to trust?–predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Front. Psychol.* 10, 2917. doi:10.3389/fpsyg.2019.02917

Kurtzberg, T. R., Naquin, C. E., and Belkin, L. Y. (2009). Humor as a relationship-building tool in online negotiations. *Int. J. Conflict Manag*. 20, 377–397. doi:10.1108/10444060910991075

Lambert, D. (2004). *Body language*. (London, UK: Harper Collins).

Lanzer, M., Babel, F., Yan, F., Zhang, B., You, F., Wang, J., et al. (2020). "Designing communication strategies of autonomous vehicles with pedestrians: an intercultural study", in 12th international conference on automotive user interfaces and interactive vehicular applications. Automotive UI '20, 10.

Lee, J. J., Knox, B., Baumann, J., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Front. Psychol*. 4, 893. doi:10.3389/fpsyg.2013.00893

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., and Rybski, P. (2010). "Gracefully mitigating breakdowns in robotic services", in 2010 5th ACM/IEEE international conference on human-robot interaction (HRI), IEEE, 203–210.

Lee, N., Kim, J., Kim, E., and Kwon, O. (2017). The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *Int. J. Soc. Robot*. 9, 727–743. doi:10.1007/s12369-017-0420-0

Lee, S. A., and Liang, Y. J. (2019). Robotic foot-in-the-door: using sequential-request persuasive strategies in human-robot interaction. *Comput. Hum. Behav*. 90, 351–356. doi:10.1007/978-981-15-5784-2_1

Lee, Y., Bae, J.-E., Kwak, S. S., and Kim, M.-S. (2011). "The effect of politeness strategy on human - robot collaborative interaction on malfunction of robot vacuum cleaner", in RSS'11 (Robotics Sci. Syst. Work. Human-Robot Interact).

Ligthart, M., and Truong, K. P. (2015). "Selecting the right robot: influence of user attitude, robot sociability and embodiment on user preferences", in 2015 24th IEEE international symposium on robot and human interactive communication, RO-MAN (IEEE), 682–687.

Maaravi, Y., Ganzach, Y., and Pazy, A. (2011). Negotiation as a form of persuasion: arguments in first offers. *J. Pers. Soc. Psychol*. 101, 245. doi:10.1037/a0023331

MacArthur, K. R., Stowers, K., and Hancock, P. (2017). "Human-robot interaction: proximity and speed−−slowly back away from the robot!" in *Advances in human factors in robots and unmanned systems*. (New York, NY: Springer), 365–374.

Martinovski, B., Traum, D., and Marsella, S. (2007). Rejection of empathy in negotiation. *Group Decis. Negot*. 16, 61–76. doi:10.1007/s10726-006-9032-z

Miller, C. H., Lane, L. T., Deatrick, L. M., Young, A. M., and Potts, K. A. (2007). Psychological reactance and promotional health messages: the effects of controlling language, lexical concreteness, and the restoration of freedom. *Hum. Commun. Res*. 33, 219–240. doi:10.1111/j.1468-2958.2007.00297.x

Miller, L., Kraus, J., Babel, F., and Baumann, M. (2020). *Interrelation of different trust layers in human-robot interaction and effects of user dispositions and state anxiety*. [Manuscript submitted for publication]

Milli, S., Hadfield-Menell, D., Dragan, A., and Russell, S. (2017). "Should robots be obedient?" in IJCAI international joint conference on artificial intelligence, 4754–4760.

Mirnig, N., Stadler, S., Stollnberger, G., Giuliani, M., and Tscheligi, M. (2016). "Robot humor: how self-irony and Schadenfreude influence people's rating of robot likability", in 2016 25th IEEE international symposium on robot and human interactive communication, RO-MAN, 166–171. doi:10.1109/ROMAN.2016.7745106

Mirnig, N., Stollnberger, G., Giuliani, M., and Tscheligi, M. (2017). Elements of humor: how humans perceive verbal and non-verbal aspects of humorous robot behavior. *ACM/IEEE Int. Conf. Human-Robot Interact*, 81, 211–212. doi:10.1145/3029798.3038337

Mnookin, R. H., Peppet, S. R., and Tulumello, A. S. (1996). The tension between empathy and assertiveness. *Negot. J*. 12, 217–230. doi:10.1007/bf02187629

Moshkina, L. (2012). "Improving request compliance through robot affect", in Proceedings of the twenty-sixth AAAI conference on artificial intelligence, 2031–2037.

Mutlu, B. (2011). Designing embodied cues for dialogue with robots. *AI Mag*. 32, 17–30. doi:10.1609/aimag.v32i4.2376

Nakamura, Y., Yoshinaga, N., Tanoue, H., Kato, S., Nakamura, S., Aoishi, K., et al. (2017). Development and evaluation of a modified brief assertiveness training for nurses in the workplace: a single-group feasibility study. *BMC Nursing*. 16, 29. doi:10.1186/s12912-017-0224-4

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., and See, S. L. (2013). Making social robots more attractive: the effects of voice pitch, humor and empathy. *Int. J. Soc. Robot*. 5, 171–191. doi:10.1007/s12369-012-0171-x

Nomura, T., Kanda, T., Suzuki, T., and Kato, K. (2008). Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Trans. Robot*. 24, 442–451.

Nomura, T., and Saeki, K. (2010). Effects of polite behaviors expressed by robots: a psychological experiment in Japan. *Int. J. Synth. Emot. (IJSE)*. 1, 38–52.

Nomura, T., Suzuki, T., Kanda, T., and Kato, K. (2006). Measurement of anxiety toward robots. *Proc. - IEEE Int. Work. Robot Hum. Interact. Commun*. 46, 372–377. doi:10.1109/ROMAN.2006.314462

Paradeda, R., Ferreira, M. J., Oliveira, R., Martinho, C., and Paiva, A. (2019). "*What makes a good robotic advisor? The role of assertiveness in human-robot interaction*", in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), LNAI*, 11876 (New York, NY: Springer), 144–154. doi:10.1007/978-3-030-35888-414

Paramasivam, S. (2007). Managing disagreement while managing not to disagree: polite disagreement in negotiation discourse. *J. Intercult. Commun. Res*. 36, 91–116. doi:10.1016/j.pragma.2012.06.011

Park, H., and Antonioni, D. (2007). Personality, reciprocity, and strength of conflict resolution strategy. *J. Res. Pers*. 30, 414. doi:10.1016/j.jrp.2006.03.003

Pfafman, T. (2017). "Assertiveness," in *Encyclopedia of personality and individual differences*. Editors V. Zeigler-Hill and T. Shackelford (Berlin, UK: Springer). doi:10.1007/978-3-319-28099-81044-1

Phansalkar, S., Edworthy, J., Hellier, E., Seger, D. L., Schedlbauer, A., Avery, A. J., et al. (2010). A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. *J. Am. Med. Inf. Assoc*. 17, 493–501. doi:10.1136/jamia.2010.005264

Preuss, M., and van der Wijst, P. (2017). A phase-specific analysis of negotiation styles. *J. Bus. Ind. Market*. 32, 505–518. doi:10.1108/JBIM-01-2016-0010

Pruitt, D. G. (1983). Strategic choice in negotiation. *Am. Behav. Sci*. 27, 167–194. doi:10.1177/000276483027002005

Pruitt, D., and Rubin, J. (1986). *Social conflict: escalation, stalemate, and resolution*.(New York, NY: Random House).

Rahim, M. A. (1983). A measure of styles of handling interpersonal conflict. *Acad. Manag. J*. 26, 368–376. doi:10.5465/255985

Rahim, M. A. (1992). "Managing conflict in organizations," in *Proc. First Int. Constr. Manag. Conf. Univ. Manchester Inst. Sci. Technol*. Editors P. Fenn and R. Gameson (New York, NY: E & F N Spon), 386–395.

Raven, B. H. (1964). *Social influence and power. Tech. Rep*. (California: University of Los Angeles).

Ray, C., Mondada, F., and Siegwart, R. (2008). What do people expect from robots? In *2008. IEEE/RSJ Int. Conf. Intell. Robot. Syst*. 46, 3816–3821. doi:10.1109/IROS.2008.4650714

Reeves, B., and Nass, C. I. (1996). *The media equation: how people treat computers, television, and new media like real people and places*. (Cambridge, UK: Cambridge University Press).

Robert, L., Alahmad, R., Esterwood, C., Kim, S., You, S., and Zhang, Q. (2020). *A review of personality in human–robot interactions*. Available at SSRN 3528496 doi:10.2139/ssrn.3528496

Rosenthal-von der Pütten, A. M., Krämer, N. C., and Herrmann, J. (2018). The effects of humanlike and robot-specific affective nonverbal behavior on perception, emotion, and behavior. *Int. J. Soc. Robot*. 10, 569–582. doi:10.1007/s12369-018-0466-7

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., and Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *Int. J. Soc. Robot*. 5, 17–34. doi:10.1007/s12369-012-0173-8

Roubroeks, M. A. J., Ham, J. R. C., and Midden, C. J. H. (2010). "The dominant robot: threatening robots cause psychological reactance, especially when they have incongruent goals", in International Conference on persuasive technology. (Heidelberg: Springer), 174–184. doi:10.1007/978-3-642-13226-118

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). "Would you trust a (faulty) robot?", in Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction - HRI'15, 141–148. doi:10.1145/2696454.2696497

Salem, M., Ziadee, M., and Sakr, M. (2013). "Effects of politeness and interaction context on perception and experience of HRI", in International conference on social robotics. (Berlin, UK: Springer), 531–541.

Sandoval, E. B., Brandstetter, J., Obaid, M., and Bartneck, C. (2016). Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. *Int. J. Soc. Robot*. 8, 303–317.

Saunderson, S., and Nejat, G. (2019). It would make me happy if you used my guess: comparing robot persuasive strategies in social human-robot interaction. *IEEE Robot. Autom. Lett*. 4, 1707–1714. doi:10.1109/LRA.2019.2897143

Savela, N., Turja, T., and Oksanen, A. (2018). Social acceptance of robots in different occupational fields: a systematic literature review. *Int. J. Soc. Robot*. 10, 493–502. doi:10.1007/s12369-017-0452-5

Shapiro, D. L., and Bies, R. J. (1994). Threats, bluffs, and disclaimers in negotiations. *Organ. Behav. Hum. Decis. Process*. 60, 14–35.

Shimada, M., Kanda, T., and Koizumi, S. (2012). "How can a social robot facilitate children's collaboration?", in International conference on social robotics. (Berlin, UK: Springer), 98–107.

Siegel, M., Breazeal, C., and Norton, M. I. (2009). "Persuasive robotics: the influence of robot gender on human behavior", in 2009 IEEE/RSJ international conference on Intelligent robots and systems, IROS, 2563–2568. doi:10.1109/IROS.2009.5354116

Sjöbergh, J., and Araki, K. (2009). "*Robots make things funnier*", in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 5447 LNAI,*. 306–313. doi:10.1007/978-3-642-00609-827

Srinivasan, V., and Takayama, L. (2016). "Help me please: robot politeness strategies for soliciting help from people", in Proceedings of the 2016 CHI conference on human factors in computing systems - CHI'16, 4945–4955. doi:10.1145/2858036.2858217

Stange, S., and Kopp, S. (2020). "Effects of a social robot's self-explanations on how humans understand and evaluate its behavior", in ACM/IEEE Int. Conf. Human-robot Interact. (Washington, D.C. IEEE Computer Society), 619–627. doi:10.1145/3319502.3374802

Strait, M., Canning, C., and Scheutz, M. (2014). "Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance", in Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction, HRI'14, 479–486. doi:10.1145/2559636.2559670

Stuhlmacher, A. F., Gillespie, T. L., and Champagne, M. V. (1998). The impact of time pressure in negotiation: a meta-analysis. *Int. J. Conflict Manag.* 9, 97–116. doi:10.1108/eb022805

Sundstrom, E., and Altman, I. (1976). Interpersonal relationships and personal space: research review and theoretical model. *Hum. Ecol.* 4, 47–67.

Sung, J., Grinter, R. E., and Christensen, H. I. (2010). Domestic robot ecology. *Int. J. Soc. Robot.* 2, 417–429. doi:10.1007/s12369-010-0065-8

Sung, J. Y., Grinter, R. E., Christensen, H. I., and Guo, L. (2008). "Housewives or technophiles? understanding domestic robot owners", in HRI 2008 - Proc. 3rd ACM/IEEE Int. Conf. Human-Robot Interact. Living with Robot, 129–136. doi:10.1145/1349822.1349840

Tay, B. T., Low, S. C., Ko, K. H., and Park, T. (2016). Types of humor that robots can play. *Comput. Human Behav.* 60, 19–28. doi:10.1016/j.chb.2016.01.042

Thacker, R. A., and Yost, C. A. (2002). Training students to become effective workplace team leaders. *Int. J. Team Perform. Manag.* 8, 89.

Thomas, J., and Vaughan, R. (2018). After you: doorway negotiation for human-robot and robot-robot interaction. IEEE international conference on intelligent robots and systems, 3387–3394. doi:10.1109/IROS.2018.8594034

Thomas, K. W. (1992). Conflict and conflict management: reflections and update. *J. Organ. Behav.* 13, 265–274. doi:10.1002/job.4030130307

Thompson, L. L., Wang, J., and Gunia, B. C. (2010). Negotiation. *Annu. Rev. Psychol.* 61, 491–515. doi:10.1146/annurev.psych.093008.100458

Thorndike, E. L. (1998). Animal intelligence: an experimental study of the associate processes in animals. *Am. Psychol.* 53, 1125.

Thunberg, S., and Ziemke, T. (2020). "Are people ready for social robots in public spaces?", in HRI 2020: Companion of the 2020 ACM/IEEE international conference on human-robot interaction. Association for Computing Machinery (ACM), 482–484. doi:10.1145/3371382.3378294

Tversky, A., and Kahneman, D. (1989). "*Rational choice and the framing of decisions*", in *Multiple criteria decision making and risk analysis using microcomputers*. (Berlin, UK: Springer), 81–126.

Van Der Laan, J. D., Heino, A., and De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transport. Res. C Emerg. Technol.* 5, 1–10.

Vollmer, A.-L. (2018). "Fears of intelligent robots", in Companion of the 2018 ACM/IEEE International conference on human-robot interaction, 273–274.

Vorauer, J. D., and Claude, S. D. (1998). Perceived versus actual transparency of goals in negotiation. *Pers. Soc. Psychol. Bull.* 24, 371–385. doi:10.1177/0146167298244004

Walters, M. L., Dautenhahn, K., Woods, S. N., Koay, K. L., Te Boekhorst, R., and Lee, D. (2006). Exploratory studies on social spaces between humans and a mechanical-looking robot. *Connect. Sci.* 18, 429–439.

Weber, K., Ritschel, H., Aslan, I., Lingenfelser, F., and André, E. (2018). "How to shape the humor of a robot - social behavior adaptation based on reinforcement learning", in Proceedings of the 20th ACM international conference on multimodal interaction, 154–162. doi:10.1145/3242969.3242976

Wilson, C. P. (1979). *Jokes: form, content, use, and function*, 16 (New York, NY: Academic Press).

Wilson, K. L., Lizzio, A. J., Whicker, L., Gallois, C., and Price, J. (2003). Effective assertive behavior in the workplace: responding to unfair criticism. *J. Appl. Soc. Psychol.* 33, 362–395.

Wisp, L. (1987). History of the concept of empathy. *Empathy Dev.* 19, 17–37.

Wullenkord, R., and Eyssel, F. (2019). Imagine how to behave: the influence of imagined contact on human–robot interaction. *Philos. Trans. R. Soc. B.* 374, doi:20180038

Wullenkord, R., Fraune, M. R., Eyssel, F., and Šabanović, S. (2016). "Getting in touch: how imagined, actual, and physical contact affect evaluations of robots", in 2016 25th IEEE international symposium on robot and human interactive communication, RO-MAN (IEEE), 980–985.

Xin, M., and Sharli, E., (2007). "*Playing games with robots - a method for evaluating human-robot interaction*", in*Human Interact, and Robot*. (Jamaica: Itech Education and Publishing), 522. doi:10.5772/5208

Xu, K., and Lombard, M. (2016). "Media are social actors: expanding the casa paradigm in the 21st century", in Annual conference of the international communication association, 1–47.

Yanco, H. A., and Drury, J. (2004). "Classifying human-robot interaction: an updated taxonomy", in 2004 IEEE Int. Conf. Syst. Man Cybern., 3, 2841–2846.

Young, J. E., Hawkins, R., Sharlin, E., and Igarashi, T. (2009). Toward acceptable domestic robots: applying insights from social psychology. *Int. J. Soc. Robot.* 1, 95.

Zhu, B., and Kaber, D. (2012). Effects of etiquette strategy on human-robot interaction in a simulated medicine delivery task. *Intell. Serv. Robot.* 5, 199–210. doi:10.1007/s11370-012-0113-3

Ziefle, M., and Valdez, A. C. (2017). "Domestic robots for homecare: a technology acceptance perspective", in International conference on human aspects of IT for the aged population, (New York, NY: Springer), 57–74.

Zuluaga, M., and Vaughan, R. (2005). "Reducing spatial interference in robot teams by local-investment aggression", in 2005 IEEE/RSJ international conference on Intelligent robots and systems. IEEE, 2798–2805.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 22 March 2021
doi: 10.3389/fpsyg.2021.640186

The Role of Frustration in Human–Robot Interaction – What Is Needed for a Successful Collaboration?

Alexandra Weidemann[1,2]* *and Nele Rußwinkel*[1]

[1] *Cognitive Modeling in Dynamic Human-Machine Systems, Faculty of Mechanical Engineering and Transport Systems, Department of Psychology and Ergonomics, Technical University of Berlin, Berlin, Germany,* [2] *Junior Research Group MTI-engAge, Control Systems Group, Department of Electrical Engineering and Computer Science, Faculty of Electrical Engineering and Computer Science, Technical University of Berlin, Berlin, Germany*

OPEN ACCESS

Edited by:
Andreas Riener,
Technische Hochschule Ingolstadt,
Germany

Reviewed by:
Elmira Yadollahi,
École Polytechnique Fédérale
de Lausanne, Switzerland
Paul Adam Bremner,
University of the West of England,
United Kingdom

***Correspondence:**
Alexandra Weidemann
a.weidemann@tu-berlin.de

Specialty section:
This article was submitted to
Emotion Science,
a section of the journal
Frontiers in Psychology

Received: 10 December 2020
Accepted: 18 February 2021
Published: 22 March 2021

Citation:
Weidemann A and Rußwinkel N
(2021) The Role of Frustration
in Human–Robot Interaction – What Is
Needed for a Successful
Collaboration?
Front. Psychol. 12:640186.
doi: 10.3389/fpsyg.2021.640186

To realize a successful and collaborative interaction between human and robots remains a big challenge. Emotional reactions of the user provide crucial information for a successful interaction. These reactions carry key factors to prevent errors and fatal bidirectional misunderstanding. In cases where human–machine interaction does not proceed as expected, negative emotions, like frustration, can arise. Therefore, it is important to identify frustration in a human–machine interaction and to investigate its impact on other influencing factors such as dominance, sense of control and task performance. This paper presents a study that investigates a close cooperative work situation between human and robot, and explore the influence frustration has on the interaction. The task for the participants was to hand over colored balls to two different robot systems (an anthropomorphic robot and a robotic arm). The robot systems had to throw the balls into appropriate baskets. The coordination between human and robot was controlled by various gestures and words by means of trial and error. Participants were divided into two groups, a frustration- (FRUST) and a no frustration- (NOFRUST) group. Frustration was induced by the behavior of the robotic systems which made errors during the ball handover. Subjective and objective methods were used. The sample size of participants was N = 30 and the study was conducted in a between-subject design. Results show clear differences in perceived frustration in the two condition groups and different behavioral interactions were shown by the participants. Furthermore, frustration has a negative influence on interaction factors such as dominance and sense of control. The study provides important information concerning the influence of frustration on human–robot interaction (HRI) for the requirements of a successful, natural, and social HRI. The results (qualitative and quantitative) are discussed in favor of how a successful und effortless interaction between human and robot can be realized and what relevant factors, like appearance of the robot and influence of frustration on sense of control, have to be regarded.

Keywords: human–robot interaction (HRI), frustration, collaboration, influence, recommendations

INTRODUCTION

Robots are no longer just tools in industrial context. Soon, robots will become part of our daily life. The vision is that robots interact with humans in close collaboration without security shelters in between. In a collaborative situation, according to Onnasch et al. (2016) humans and robots work on common goals and subgoals, which are assigned according to the situation during the collaboration and take place in the same workspace.

The challenge for human–robot interaction (HRI) research is to design a successful and enjoyable interaction. The identification and measurement of factors that play a relevant role in successful collaborations is crucial regarding the design and development of a suitable robot system and the direct interaction. If the robot does not meet the requirements, needs and perspectives of the user, or if those are not taken into account, the robot will most probably not be accepted by the user (Davis, 1989; Venkatesh et al., 2003; Heerink et al., 2007; Broadbent et al., 2012; Smarr et al., 2014). Various lines of research (such as Riek et al., 2009; Waytz et al., 2010; Salem et al., 2015; Abd et al., 2017; Ciardo et al., 2018; Onnasch and Roesler, 2019) investigated different aspects like trust, appearance, anthropomorphism, and acceptance that play a role in HRI. An important aspect of human-centered research in HRI are human emotions during the interaction, especially negative emotions. One negative emotion that is often mentioned in dealing with technology, is frustration (Ceaparu et al., 2004; Lazar et al., 2006). Frustration arises when a person has the expectation to reach a goal but still fails to achieve it after repetitive attempts (based on Freud, 1921; Russell, 1980; Amsel, 1992; Scherer, 2005; Bortz and Doering, 2013).

Expectations

Humans have specific expectations regarding the details of the interaction with a robotic system based on, e.g., the appearance of the robot system, the way of conducting the task with the system often relating to the similarity to human–human interaction (HHI), like the way of communication (verbal and non-verbal) and social behavior toward the interaction partner and social norms (Compagna et al., 2016; Beer et al., 2017; Jerčić et al., 2018). Humans use HHI mechanisms, like proxemic behavior, interpretation of the other's intention, the way of communication, and social, physical, behavioral cues, to perceive robots as autonomous social agents, as socially present human employees (Fiore et al., 2013). It has been shown that humans treat computers as teammates with personality (Nass et al., 1995, 1996). Humans tend to behave socially not only toward other humans but also toward robots (Reeves and Nass, 1998; Dautenhahn, 2007). Without prior training, humans prefer natural and intuitive communication in use with the technical system (Dautenhahn et al., 2005). There is a tendency for people to prefer human-like attributes in robots (Kiesler and Hinds, 2004; Walters et al., 2008). It has been shown that humans were better able to empathize with this type of robot (Riek et al., 2009) and this assumingly leads people to ascribe human-like mental abilities to the robot (e.g., intentions, emotions, cognition) (Waytz et al., 2010; Schneider, 2011).

Regarding the question on how to realize successful collaborative working situations, it is helpful to analyze human–human collaboration situation. Humans have developed a number of abilities to achieve joint action (Sebanz et al., 2006). Mechanisms such as joint attention and other cognitive mechanisms for sharing representations of objects and events as well as common task knowledge help us to initiate and coordinate joint action. Whenever actions of the partner indicate a mismatch of the representation of the common goal and the way of how to achieve this goal, an immediate facial expression follows and informs the partner without too much explicit communication. Therefore, such emotional facial reactions could also be a very relevant indicator for a successful human-robot collaboration.

To evaluate human reactions to different kinds of robots with varying outer appearance, many studies have used pictures or videos (e.g., Bartneck et al., 2007). However, two-dimensional images cannot represent the complex three-dimensional appearance, movement and sounds of social HRI (e.g., Wainer et al., 2007). Therefore, it is important for studies investigating HRI, to use at least two different kind of robots (with differences in human-like appearance) to prevent a misinterpretation of behavior and considering a broader variability of reactions to different robotic systems.

For these reasons it is interesting to consider the appearance of the robot, expectations that arise and to draw comparisons to HHI for designing robot systems and HRI.

Negative Emotion – Frustration

If the expectations of a human partner on the robot are disappointed or not fulfilled, negative emotions like frustration can arise and even lead to the termination of the interaction. Emotions can occur during all kind of actions and mental operations (Picard, 1997), they motivate actions and have influence on performance, trust, and acceptance during an interaction and on the interaction behavior itself (Brave and Nass, 2002).

The emotional experience of frustration can be caused by simple events such as time delays and errors that can occur due to lack of knowledge and insufficient training in human–computer interaction (HCI) (Bessière et al., 2004; Lazar et al., 2006). Working with a computer agent that the user does not trust leads to the development of frustration (Hirshfield et al., 2011). Examples in the literature of frustration in HRIs are situations such as e.g., behavioral errors by the robot like dropping a bottle or moving to the wrong takeover-location in a bottle handover task with the robot (Abd et al., 2017). In interactive situations with different robots, the participants are more frustrated by such kinds of technical failures than when experiencing a social norm violation, for example "not looking directly at the person it is talking to" (Giuliani et al., 2015, p. 3) (Giuliani et al., 2015). For these reasons, technical failures were used in our study (see also section "Experimental Description"). Such technical failures could be used in studies to intentionally induce frustration to participants in such interaction situations to generate a perceived increase in frustration. In such cases humans usually show immediate emotional feedback to the robot in form of reactions such as facial expressions (Lang et al., 2010).

Frustration leads to lower task productivity (Waterhouse and Child, 1953; Klein et al., 2002; Powers et al., 2011), slower response times (Chen et al., 1981), longer decision-making time (Lerner et al., 2015), prolonging content acquisition on learning (Amsel, 1992), and lower learning efficiency (Kort et al., 2001; Graesser et al., 2005; Woolf et al., 2009). Decreased motivation (Weiner, 1985), user satisfaction, and lacking trust (Lazar et al., 2006; Hirshfield et al., 2011) are evoked by frustration. It was found that frustration triggers a rise in arousal, which enhances cognitive performance, and is associated to high workload (e.g., Whinghter et al., 2008). Therefore, whenever the perception of frustration could be prevented, this would cause a benefit on the further interactive process and the quality of the task conductance.

Various authors found a direct influence on the acceptance of a technical system and trust on the decrease of frustration (Giuliani et al., 2015; Yang, 2016; Abd et al., 2017). It was found that the sense of dominance was low when frustration was high in a task with high attentional demands (Weidemann and Rußwinkel, 2019). In this study, dominance was viewed and questioned as control and the ability of being in control of a situation. The concepts of dominance and control in the study described in this paper were considered separately by extending the SAM questionnaire (for more details see section "Questionnaires"). The dominance dimension in the SAM questionnaire represents changes in degree of control, the maximum control in the situation is presented by a large figure (Bradley and Lang, 1994). In this study, "dominance" is defined as superiority in interaction and also over the interaction partner and "control" as control in the situation, over one's own action and through action, i.e., also as the difference between the perception of an event in a situation and the intended effect (Pacherie, 2007; Haggard and Chambon, 2012). Dominance is an important factor for the judgment of the interaction, partner and communication in a social interaction (Ng and Bradac, 1993; Berger, 1994). The importance of dominance has also been shown in the results of the SAM questionnaire in our past study on frustration.

The two terms sense of control and sense of agency are connected in psychology. Sense of agency refers to "being in control both of one's own actions and through them" (Haggard and Tsakiris, 2009, p. 242). Being able to realize intended actions and the expected outcome with the robot would therefore result in a higher sense of control. Such a factor is interesting in regard to how successful a tool is used for a certain aim as well as how successful I am in an interaction with another person e.g., "am I successful in order to make myself understood by the other person," or in other words, "do I experience the intended effect that I tried to cause by my actions?" Ciardo et al. (2018) suggest that sense of agency is negatively affected by frustration in the interaction with an embodied robot similarly, to interacting with other humans.

In that sense repeated unsuccessful HRIs related to a chosen aim leads to perceived frustration of the human partner. The identification of such unsuccessful frustrating events would enable the implementation of solution functions, e.g., for the HHI.

As can be seen, it is important to be able to identify and minimize frustration. Emotions are object-directed and have a characteristic experience and the occurrence of physiological changes and behavioral patterns is evident (Klug, 2012). In the literature several methods are reported to access emotions, these can be divided into subjective (like questionnaires) and objective (like psychophysiological methods) methods.

How to Measure Emotions

Subjective measures of emotions such as self-report methods are efficient and easy to administer, they are beneficial to determine emotions. However, participants are susceptible to time effects and may respond based on social desirability (Mauss and Robinson, 2009; Lopatovska and Arapakis, 2011) or have no direct access to the emotional experience.

During the experience of emotions specific physiological changes occur in the human body (Peterson et al., 2015). Because the measurements of such physiological changes can be recorded parallel to the occurrence of the emotion in contrast to subjective methods. An additional use of psychophysiological methods would support the determination of emotions. Vyzas and Picard have shown correlation between various emotions (such as joy and frustration) and physiological signals (like pulse and galvanic skin response) (Vyzas and Picard, 1999). On the downside, physiological measurements are ambiguous, and the best methodological combination of measurements remains to be found especially regarding different experimental settings.

The multicomponent phenomenon frustration often occurs during human–machine interaction (Ceaparu et al., 2004; Lazar et al., 2006) and initiates not only changes in facial expression, but also in posture, physiology, or behavior (Scherer, 2005). It was found that heart rate variations are sensitive to frustration and the heart rate itself is positively correlated with this emotion (Wulfert et al., 2005; Washington and Adviser-Jones, 2011; Yuan et al., 2014). During incorrectly completed tasks, facial muscle activity may also provide evidence of frustration (Jost, 1941; Hamm et al., 2011; Hazlett, 2013; Gao et al., 2014; Lerner et al., 2015). But all these findings are not robust enough to be used in isolation to measure frustration. Therefore, a multi-method approach to measuring frustration is used in this study.

Aim

It seems that the emotional experience of frustration and its influence on interaction factors, and interaction quality could provide a good guideline for the evaluation of robot systems, and for the recommendation of the design of a pleasant and successful HRI.

To gain a deeper understanding of these possibilities, we follow one main question in this paper:

How does frustration influence HRIs?

To investigate this question a human–robot collaborative experiment was designed, consisting of a task with a common goal including handover scenarios. The participants interacted with two different robot systems, investigate the range of changes in behavior due to the technical system used. In the experiment different measurements of frustration were applied, which have been used before in similar studies. One aim of the study was to induce and measure frustration, among others with questionnaires. The second aim was to investigate the influence frustration has on the HRI.

Hypotheses

Based on findings from related work on frustration and robot appearance in psychology, HCI and HRI, we developed four hypotheses for the study:

H1: Technical errors by the robot lead to perceived frustration by the participants.
H2: Frustration leads to decreased dominance, sense of control, and self-reported performance.
H3: Frustration leads to lower rating regarding acceptance of the robot systems.
H4: The interaction with the more human-like robot (here "*Pepper*") is preferred, among other aspects due to the human-like appearance and similarity to HHI. This leads to an attribution of human-like abilities to the robot and to a tendency to forgive mistakes, in contrast to a more technical looking robot that would be expected to behave more precise.

MATERIALS AND METHODS

Experimental Motivation

A collaboration task was chosen to investigate direct interactions, since the human shares the workspace with the robot to perform the common task. In this study the interaction corresponds to a task processing (in the following called interaction task).

A good example of a close interaction task is a handover scenario with a robotic system. Since different colored balls needed to be handed over from the human to the robot and be placed in specific baskets relevant components such as feedback (robot and human), joint actions and giving and perceiving instructions were relevant for the quality of task completion. Similar scenarios have been investigated elsewhere (Rasmussen, 1982; Giuliani et al., 2015; Abd et al., 2017; Honig and Oron-Gilad, 2018) with differing research questions. For the scenario in the present example, technical execution failures of the robot were initiated, like dropping the ball, to induce frustration.

A task with a common goal is helpful for the development of negative emotions, such as frustration. After all, not achieving a common goal that is relevant to you because your partner fails can lead to frustration.

Two different robot systems (**Figures 1A,B**) were used taking into consideration that the appearance, movement could form different expectations and might have a strong influence on the interactive behavior of the participant and on the evaluation of the interaction task. For a systematic investigation of such kind of influences a broader variety of robotic systems would have been necessary. In other studies, usually only one type of robot is investigated. In the study described here, a person is working on the same task interacts with two systems (one after the other), so they can (be) compared directly. The requirements of the two to be chosen robotic systems were (1) the ability to physically interact with the participant (at a similar paste) and (2) to find two systems that differ in humanoid appearance, such as a social and industrial robot (Chanseau, 2019).

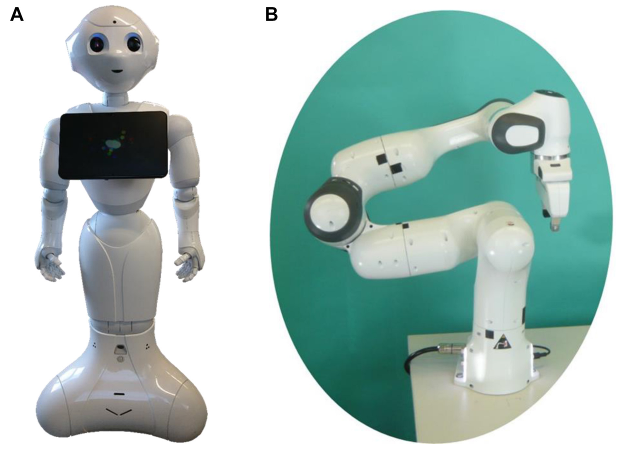


FIGURE 1 | (A) On the left the robot "*Pepper*" from Aldebaran Robotics SAS and SoftBank Mobile Corp and **(B)** On the right the robot "*Panda*" from FRANKA EMIKA. The photos of the robots were taken and edited by Alexandra Weidemann.

The methods (questionnaires and interviews) used have already proven in other studies to determine emotions or even frustration. In addition, these methods have been investigated based on a multimodal approach in order to investigate which methods are best suited to measure frustration in HRIs.

Questionnaires, video recordings (to counterbalance the self-assessment problem (Bethel and Murphy, 2010) and to evaluate reactive behavior showed by participants) and interviews (to provide further insights into the participants state of mind) are frequently applied as methods in the observation of interactions in various studies and were also used here (Chanseau, 2019).

Feedback given by the robot, in our case status of the system (open for instructions or not), is very important for good communication between two parties in an interaction. Here the chosen feedback channel was visual and realized as LED-feedback, which has been shown to be useful for example in a study by MTI-engAge project.

Experimental Description

Study Design and Participants

The HRI study was done in a between-subject design with 30 healthy participants [age: 18–35 years; N (male) = 14, N (female) = 16]. The average age was 29.1 (SD = 5.2). Subjects were recruited via notices at universities in Berlin and the subject portal of the Technical University of Berlin. The subjects were randomly divided into two condition groups: frustration (FRUST) and no frustration (NOFRUST) which was considered as independent variable.

Technical Systems

Robotic systems

The subjects interacted with two different robot systems, a humanoid robot ("*Pepper*" from Aldebaran Robotics SAS and SoftBank Mobile Corp) and a robotic arm ("*Panda*" from FRANKA EMIKA). The robots were controlled by a Wizard-of-Oz scenario (controlled by a specially written computer program), so the experimenter generated the reactions of the robots during the interaction tasks for practical and safety reasons.

LED-feedback

To enable the robot to give feedback in response to a "trigger input" from the subject an LED-feedback was developed. The robots gave feedback to the human about their current "state" via three colors of a LED lamp. If the LED was "green," instructions (with the help of gestures and/or words) could be given to the robot. If the LED was "orange," the robot "processed" the input from human. If the LED turned "red," then the robot either did not understand the input or the input was incorrect.

Experimental Setup

The interaction tasks (one with "*Pepper*" and one with "*Panda*") took place in rooms separated by curtains, so that the subjects were "alone" with the robot and visually shielded from other people (see **Figure 2**). Each interaction-place was divided into two areas: the area for the human (green area) and the area of the robot (red area), which the human was not allowed to enter with any part of the body. The subject changed stations during the experiment. At station 1, the subject filled out the questionnaires before and after the interaction tasks. The interaction tasks took place at station 2. The Wizard of Oz's (the person that controlled the robot) seat was at robot height and hidden behind the curtains. From there, the wizard was able to observe the participants with the help of cameras above the station 2, and controlled the robot.

Experimental Procedure

The procedure of the experiment was divided into three blocks (**Figure 3**). In the first block, general questionnaires (pre-testing) were filled out. The interaction tasks with a robot system (at station 2) and the completion of the corresponding questionnaires took place in block 2. Thus, the participants performed block 2 twice. The interaction tasks with the two robot systems took place successively in randomized order. This served to avoid a sequence effect. Questionnaires about the health of the human interaction partner, the knowledge about the triggers, the robot system, and the interaction were given at four different time points throughout the block 2, before each interaction task (T1 and T3) and after each interaction task (T2 and T4) (at station 1). In block 3 final questionnaires regarding both interaction tasks were filled out and optionally an interview was performed.

Each interaction task in block 2 included 11 trials, since a maximum of 11 balls should be handed over (handover scenarios). Within a trial, no errors or two to three errors could occur. In the FRUST-group, errors occurred in nine trials (trial 2, 3, 4, 5, 6, 8, 9, 10, and 11). In 2 trials (trial 4 and 7) errors occurred in the NOFRUST-group. The experimenter determined, according to this rules, in which trial errors arose before the study started.

Block 2: interaction task

The different handover situations with the robots were controlled by a Wizard-of-Oz scenario (controlled by a specially written computer program and the experimenter). These are handover scenarios in which the subject should give colored balls (yellow and blue) to the robot and, with the help of gestures and words, get the robot to throw the ball in a corresponding colored basket in the room separated from the human.

The participants had two subtasks. In the first subtask, at least three balls of each color had to be placed with the help of the robot into the corresponding basket. In the second subtask the participants had to find out which gestures and words, so-called triggers, caused the robot to react and release the ball into the basket. The type of trigger words (color, direction) and trigger gestures (pointing gesture, color card) were known by the participants, but not which robot reacted to which corresponding trigger (word or gesture) or trigger combination (word and gesture) with the desired reaction (release of the ball into the corresponding basket). The participants stated their knowledge about the triggers in the knowledge inquiry at the end of the experiment (see also section "Questionnaires").

The interaction tasks were divided into four different phases:

(1) attracting
(2) handing over the ball
(3) choosing the trigger
(4) the robot's passing of the ball.

In the attracting phase, the subjects should attract the robot, for example by calling over, so it would moves toward the human to receive the ball with the robot's gripper for the ball transfer. After the ball was successfully handed over, the robot moved to a so-called "waiting position" and the subject could select the trigger to find out how the robot reacts to the trigger. This was also supported by the LED-feedback. If the trigger was selected correctly, the robot released the ball into the corresponding basket in its area.

The technical errors caused by the robot occurred during the ball handover phase of the four interaction task phases. There were four different types of technical errors:

(1) the gripper remained open
(2) the gripper remained closed
(3) the gripper picked up the ball and dropped it in the area of the human
(4) the gripper picked up the ball and dropped it in its area.

There were more errors in the FRUST-group than in the NOFRUST-group, so the subjects in the FRUST-group were supposed to experience frustration.

During the interaction tasks, video recordings (from the front and from the side, see also **Figure 2**) were made. Short interviews were conducted with a certain number of subjects about the interactions.

Questionnaires

All questionnaires were filled out on the computer.

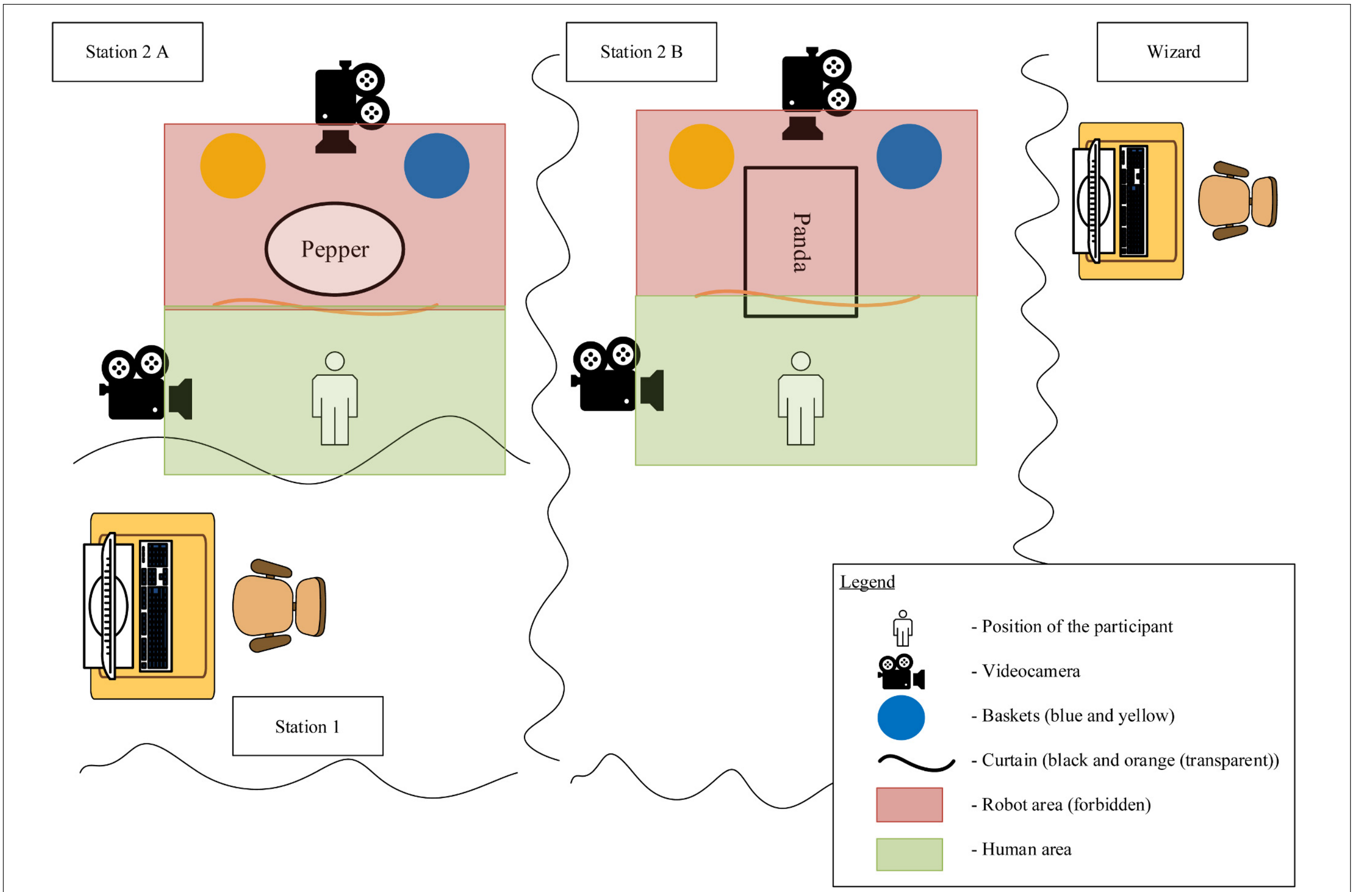


FIGURE 2 | Setup of the human–robot interaction experiment from above.

The pre-testing phase in block 1 included questionnaires on the affinity for technology, general well-being, and emotion regulation.

The following described questionnaires expect the post-post study questionnaire were given at four different points in time throughout the experiment, before each interaction task (T1 and T3) and after each interaction task (T2 and T4).

The three following questionnaires have to be filled out before (T1 and T3) and after the interaction task (T2 and T4) (**Figure 3**, see block 2). A 6-scale questionnaire about different emotions (like satisfaction and frustration) and condition (like tiredness) of the human (EaCQ) was based on Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988; Krohne et al., 1996) and BSKE21 (Janke et al., 1988, 1995; Janke and Debus, 2003). This questionnaire and the self-assessment manikin (SAM) (Bradley and Lang, 1994) ranged from 1 to 6. SAM and EaCQ were performed to be able to evaluate the emotional state over the task period. The third questionnaire was the NASA‘s Task Load Index (NASA- TLX) (Hart and Staveland, 1988), which was used to determine task performance and frustration. The scale was converted linearly into percentage scales. These questionnaires were already used in other literature to identify changes in emotions, especially frustration (for example Yuan et al., 2014; Ihme et al., 2016, 2017, 2018).

The SAM questionnaire (Bradley and Lang, 1994) was extended by a "control" scale. The already existing scale of dominance ranges from inferior to superior. The term "control" is supplemented in the questionnaire by the words "control of the situation."

In the knowledge inquiry, the subjects were asked about their knowledge of the trigger words or gestures acquired in the interaction and the corresponding reactions of the robots (**Figure 3**, see Block 2 T2 and T4).

In the adapted Post-Study System Usability Questionnaire (7-point scale, 1 to 6 and "specification not possible") (Lewis, 1992, 2002; Sauro and Lewis, 2012) and the adapted Godspeed questionnaire (question pairs) (Bartneck et al., 2009) the interaction tasks and the robots were evaluated (see Block 3 in **Figure 3**).

The post–post study questionnaire was used to find out which interactions were perceived as more pleasant and how subjects define frustration since several different emotions might relate individually to this emotion (such as hate, sadness, and others).

Protocol of the Wizard-of-Oz

The Wizard-of-Oz (WoO) indicated before the start of the experiment whether the participant was in the FRUST- or NOFRUST-group stating accordingly in the

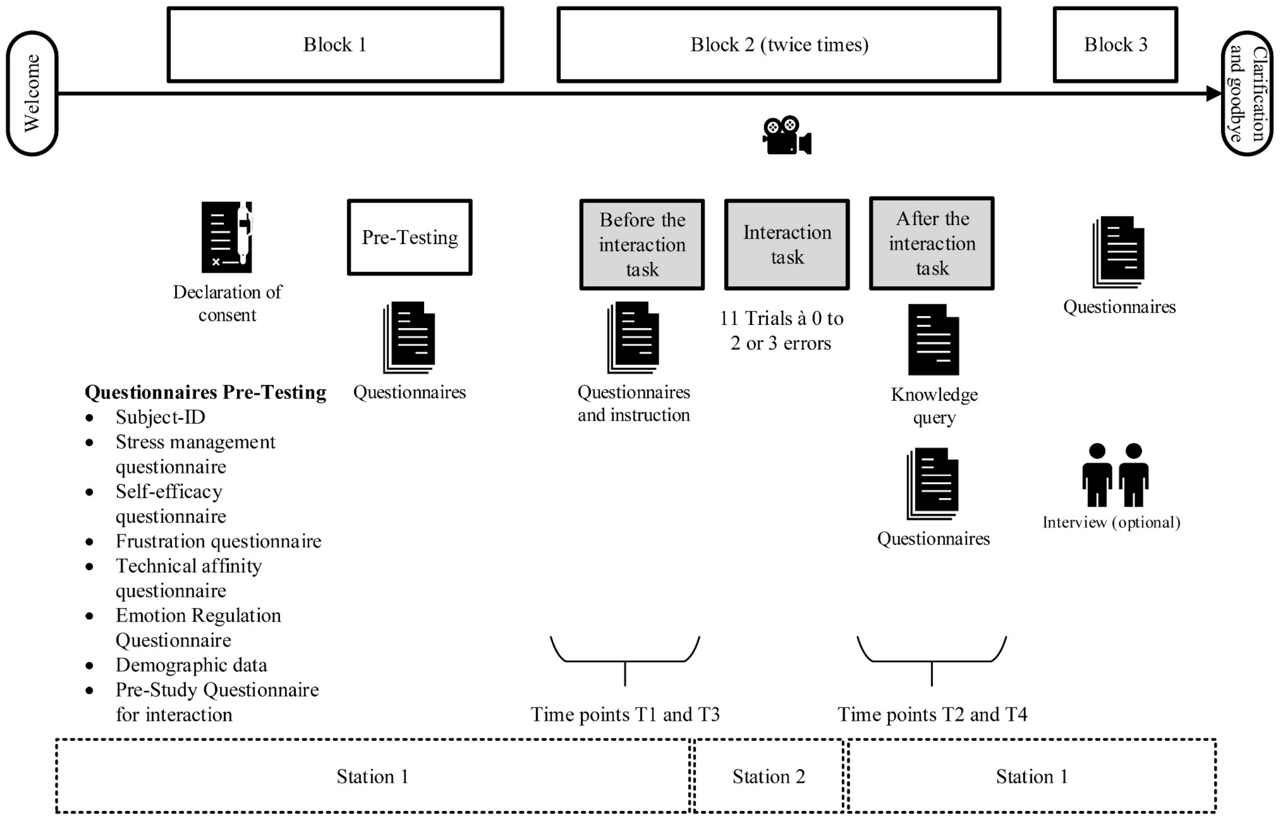

Welcome
Block 1
Block 2 (twice times)
Block 3
Clarification and goodbye
Declaration of consent
Pre-Testing
Questionnaires
Before the interaction task
Questionnaires and instruction
Interaction task
11 Trials à 0 to 2 or 3 errors
After the interaction task
Knowledge query
Questionnaires
Questionnaires
Interview (optional)
Questionnaires Pre-Testing
Subject-ID
Stress management questionnaire
Self-efficacy questionnaire
Frustration questionnaire
Technical affinity questionnaire
Emotion Regulation Questionnaire
Demographic data
Pre-Study Questionnaire for interaction
Time points T1 and T3
Time points T2 and T4
Station 1
Station 2
Station 1

FIGURE 3 | Procedure of the human–robot interaction experiment.

program of the robot: Should the subject be frustrated? Yes (key "y") or no (key "n"). This selected the appropriate program in which it was already determined in which trial which errors would occur. So the errors were not selected during the interaction task by WoO, they were already predefined.

The robot "waked up" and moved to the initial position. This movement was not seen by the participants. The WoO saw the interaction task with the help of a camera placed above the participant and the robot.

The action of the WoO within a trial could be divided in three phases:

(1) Activation of the movement to the handover position
(2) Action after an error or no error answering the question, if the participant choose the right trigger (gesture or trigger)
(2a) in case of an error: the answer was "no." A new ball transfer was possible. It started again with phase 1.
(2b) in case of no error: the answer was "yes" after choosing the right trigger and "no" after choosing the wrong trigger.
(3) Transfer the ball to the corresponding container after the right trigger. After the release of the ball, the next trial started with phase 1.

In the following the phases were explained in more detail:

(1) The movement to the handover position to pick up the ball was activated by pressing the button "t" on the keyboard after the participant called over the robot. Either the error occurred, or the handover succeeded. The robot moved to the waiting position.
(2a) In case of an error, the wizard indicated that the input of the trigger was wrong. The participant could call over again. The wizard answered the next question: Did the participant ask for a new ball transfer? y/n. The trial started again with phase 1.
(2b) In case of no error, the wizard indicated whether the input of the trigger (gesture or word or combination) was correct or incorrect: Was the input correct? y/n.
(2b1) if yes, the LED lighted green. The wizard answered to the next question for the direction of the ball release into the corresponding basket. The gripper released the ball according into the container on the right or left side of the robot.
(2b2) if no, the LED lighted red after answering the next question with "no" by the wizard: "Did the participant ask for a new ball transfer?" So the participant could test another trigger/trigger combination until the answer to the trigger choice question was "yes." Than the gripper

released the ball into the corresponding container on the right or left side of the robot.

(3) After the right trigger choice and the releasing of the ball the robot moved back to the starting position. The next trial started with phase 1.

Ethics Approval Statement

The experiment received a positive ethical vote from the ethics committee of the Technical University of Berlin.

Data Analysis

The statistical analyses were carried out using SPSS 22 IBM Corp. (2013). For analysis and in order to provide a clearer understanding of how reliable and "stable" the results are, 95% confidence interval (CI), effect size (ES) *r* (Cohen, 1988), and *p*-values were determined (Cumming, 2014). Small effect is $r = 0.1$, medium effect is $r = 0.3$, and large effect is $r = 0.5$ (Cohen, 1988; Gignac and Szodorai, 2016). Self-performance is a score of the NASA-TLX scale. *T*-tests and bivariate correlations were conducted.

The condition (FRUST or NOFRUST) is the independent variable.

RESULTS

The results are presented in several sections. The first section deals with the detectability of frustration and the definition of the term. Then the results about the behavioral reactions after the technical execution error in the video data follows. Finally, the influence of frustration on interaction factors and the evaluation of interaction and robot systems is presented. More details about the results are shown in tables (**Tables 1–5**). Each table contains columns of the time points and the factors that were considered, of the confidence intervals (lower and upper bound), the effect size *r* and the *p*-values.

Square brackets in the text signal a 95% CI, lower and upper bound. The effect size is *r*.

Before (T1 and T3) and after (T2 and T4) the respective interaction task with the robot, the participants completed questionnaires (SAM, EaCQ, and NASA-TLX) about their own perception (see also **Figure 3**, Block 2).

TABLE 1 | The table shows the results of the frustration scales of the NASA-TLX and the EaCQ (see also section "Frustration can be determined with subjective methods").

Time point	Factors	Upper bound	Lower bound	Effect size	*p*-value
After first interaction	Frustration (EaCQ 1)	0.04	1.68	0.383	0.04
	Frustration (NASA 1)	14.84	49.52	0.616	0.001
After second interaction	Frustration (EaCQ 2)	0.43	1.76	0.541	0.002
	Frustration (NASA 2)	16.48	48	0.661	0.0003

TABLE 2 | The table shows the results of the reaction of the participants after an error of the robot (see also section "Specific reactions after an error by the robot").

Time point	Factors	Upper bound	Lower bound	Effect size	*p*-value
First interaction	Smile	0.13	0.82	0.482	0.008
	Laugh	0.17	0.82	0.507	0.004
	Facial expression overall	0.89	2.34	0.655	0.0001
Second interaction	Lick one's lips	0.01	0.46	0.485	0.041
	Laugh	0.10	0.76	0.454	0.012
	Facial expression overall	0.61	2.47	0.548	0.002
FRUST first interaction	Body overall	–1.82	–0.43	0.822	0.007
FRUST second interaction	Lick one's lips	–0.85	–0.04	0.667	0.035
	Cock one's head	–0.85	–0.04	0.667	0.035

TABLE 3A | The table shows the results of the interaction factors scales of the SAM, the EaCQ, and the NASA-TLX (see also section "Dominance and sense of control differs between condition groups").

Time point	Factors	Upper bound	Lower bound	Effect size	*p*-value
After first interaction (T2)	Control (SAM)	–1.84	–0.15	0.444	0.023
After second interaction (T4)	Control (SAM)	–2. 29	–0.74	0.622	0.0005
Change during first interaction	Dominance1 (SAM)	–1.4	–0.12	0.438	0.022
Change during first interaction	Dominance2 (SAM)	–1.35	–0. 17	0.490	0.014
	Control2 (SAM)	–2.39	–0.37	0.495	0.009

Frustration Can Be Determined With Subjective Methods

The frustration score of both questionnaires (NASA-TLX and EaCQ) was higher in the FRUST-group than in the NOFRUST-group after both interaction tasks (T2 and T4) (EaCQ 1: $MD = 0.86$ [0.04, 1.68], $r = 0.383$; NASA 1: $MD = 32.18$ [14.84, 49.52], $r = 0.616$; EaCQ 2: $MD = 1.1$ [0.43, 1.76], $r = 0.541$; NASA 2: $MD = 32.24$ [16.48, 48], r = 0.661) (**Figure 4** and for more details see **Table 1**).

Participants were statistically not significant more frustrated in the interaction tasks with the robot "*Panda*" than with the robot "*Pepper*" (first interaction task: $MD = 0.941$ [–22.78, 21.97], r = 0,007; second interaction task: MD = –6.27 [–27.3, 14.77], r = 0,125). Moreover, there is no statistically significant difference whether participants

TABLE 3B | The table shows the correlation between frustration and interaction factors (see also section "Frustration correlated negative with dominance, control and self-confidence").

Time point	Factors	Upper bound	Lower bound	Effect size	*p*-value
After first interaction (T2)	Frustration and arousal	0.295	0.789	0.578	0.001
	Frustration and dominance	–0.685	–0.128	–0.459	0.011
	Frustration and control	–0.779	–0.410	–0.601	0.0005
	Frustration and self-confidence	-0.754	-0.322	-0.576	0.001
	Frustration and eye-rolling	0.212	0.611	0.371	0.044
	Frustration and facial expression overall	0.157	0.707	0.476	0.008
	Frustration and mouth twisting	0.025	0.692	0.4	0.028
After second interaction (T4)	Frustration and arousal	0.140	0.842	0.562	0.001
	Frustration and dominance	–0.690	–0.076	–0.445	0.014
	Frustration and control	–0.842	–0.440	–0.673	0.000047
	Frustration and self-confidence	–0.858	–0.530	–0.717	0.000008
	Frustration and self-reported task performance	–0.844	–0.228	–0.587	0.001
	Frustration and head-shaking	–0.021	0.788	0.462	0.01
	Frustration and lips linking	–0.007	0.705	0.412	0.024
	Frustration and eyebrow pull together	–0.008	0.146	0.502	0.006
	Frustration and facial expression overall	0.010	0.125	0.451	0.005
	Frustration and breathing out	0.0002	0.131	0.490	0.012

TABLE 4 | The table shows the results of the reaction of the participants after an error of the robot for both robots in comparison between the condition groups (see also section "Specific reactions after an error by the robot").

		Pepper				***Panda***			
Time point	**Factors**	**Upper bound**	**Lower bound**	**Effect size**	***p*-value**	**Upper bound**	**Lower bound**	**Effect size**	***p*-value**
First interaction	Laugh	0.003	1.11	0.641	0.049				
	Smile					0.098	1.08	0.627	0.023
	Facial expression overall					1.24	3.73	0.807	0.001
Second interaction	Laugh	0.19	1.06	0.791	0.011				
	Facial expression overall	0.34	2.38	0.626	0.013				
	Speech overall	0.01	1.49	0.671	0.048				
	Lick one's lips					0.039	0–85	0.667	0.035

first interacted with "*Pepper*" or with "*Panda*" in both conditions (first interaction: FRUST: *MD* = –6.54 [–38.42, 25.34], *r* = 0.115; NOFRUST: *MD* = 2.67 [–17.65, 22.98], *r* = 0.111; second interaction: FRUST: *MD* = –15.5 [–43.92, 12.91], *r* = 0.291; NOFRUST: *MD* = 10.76 [–2.86, 24.85], *r* = 0.516).

Understanding of Term Frustration by Participants

In a free text field, the participants described what they understood by the word "frustration." The participants also indicated which terms (terms from the NASA-TLX and which they themselves specified) they associate to what percentage with

TABLE 5 | The table shows the results of the robot rating for both robots in comparison between the condition groups for each interaction (see also section "Robot Rating: Robots were evaluated different in condition groups").

		Pepper				***Panda***			
Time point	**Factors**	**Upper bound**	**Lower bound**	**Effect size**	***p*-value**	**Upper bound**	**Lower bound**	**Effect size**	***p*-value**
First interaction	Easy to use	–3.03	–0.74	0.703	0.003	–2.48	–0.31	0.712	0.018
	Correction of errors	–3.49	–0.95	0.726	0.002				
	Easy to brief	–2.86	–0.36	0.622	0.016				
	Good task					–2.87	–0.13	0.636	0.036
	Pleasant use					–2.19	–0.028	0.524	0.045
	productivity					–2.34	–0.41	0.649	0.009
	Clarity of reactions					–2.97	–0.28	0.611	0.022
Second interaction	Easy to use					–3.67	–1.66	0.863	0.000099
	Good task	–2.14	–0.18	0.583	0.024	–3.31	–0.57	0.684	0.010
	Pleasant use	–2.02	–0.63	0.755	0.001	–3.38	–0.74	0.738	0.006
	Productivity	–2.43	–0.21	0.652	0.025	–4.03	–0.63	0.718	0.013
	Satisfaction	–2.46	–0.04	0.527	0.044	–3.49	–0.39	0.749	0.021
	Clarity of reaction	–2.94	–0.16	0.557	0.031				
	Easy to brief	–2.24	–0.04	0.529	0.043	–3.01	–1.10	0.802	0.000499
	learning to use					–2.87	–0.47	0.687	0.011
	Overall evaluation					–4.32	–0.24	0.729	0.034

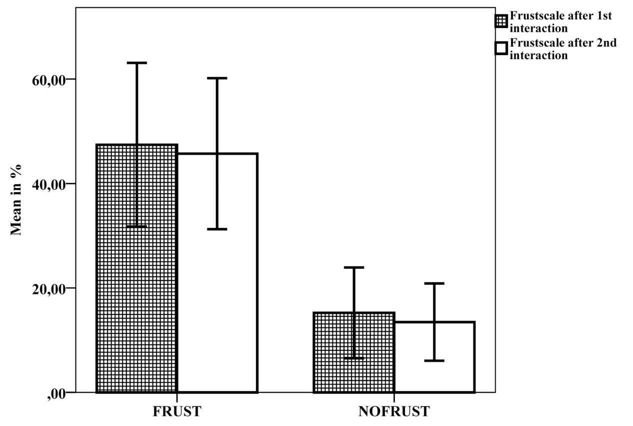


FIGURE 4 | Results of the frustration scales of the NASA-TLX questionnaire after the first (T2) and second (T4) interaction task, mean values and 95% confidence intervals.

the term frustration. In the free definitions, the participants mainly indicated "disappointed expectations" and "not reaching a goal despite repeated attempts." The term "annoyance" was given a high percentage, followed by "stress." "Irritation" and "discouragement" were in average associated to frustration to more than 50%. Other terms frequently mentioned by the participants were "helplessness" and "disappointment."

Specific Reactions After an Error by the Robot

The videos of the HRIs were scanned for reactions of the participants to the errors of the robots. Then the frequencies of the reactions were counted, i.e., it was looked whether the reaction occurred at all in the interaction task and not how often in an interaction task. In addition, the reactions were summarized in four parameter groups: gestures, facial expressions, speech and body.

The results show that mainly facial expression are shown and in this parameter group, surprisingly, mainly laughter and smiles were found. There are mainly differences in the condition groups for these reactions (for more details see **Tables 2**, **4**). Smiling and laughing was a frequent reaction after the occurrence of an error especially in the FRUST-group (see **Table 2**).

Frustration correlated positively with various reactions that participants exhibited following the robot's errors in both interaction tasks (see also **Tables 3A**,**B**). In the first interaction task, there were positive correlations between frustration and facial expressions (r = 0.476 [0.157, 0.707]), such as eye-rolling (r = 0.371 [0.212, 0.611]) and mouth-twisting (r = 0.4 [0.025, 0.692]). In the second interaction task, there were also positive correlations between facial reactions (r = 0.502 [0.175, 0.737]), such as licking lips or pulling eyebrows together (r = 0.490 [0.232, 0.734]) and frustration. In addition, there were positive correlations between frustration and head shaking and breathing out.

Dominance and Sense of Control Differs Between Condition Groups

Differences in control perception (SAM) between condition groups after the 1st (T2) and 2nd (T4) interaction task were found (SAM T2: MD = –0.995 [–1.84, –0.15], r = 0.444; SAM T4: MD = –1.51 [–2.29, –0.74], r = 0.622) (for more details see **Table 3A**).

There were differences between the groups for the factors of the SAM questionnaire items dominance and control before (T1 and T3) and after (T2 and T4) an interaction task (dominance T2: *MD* = –0.77 [–1.4, –0.12] *r* = 0.438; dominanceT4: *MD* = –0.76 [–1.35, –0.17], *r* = 0.490; control T4: *MD* = –1.38 [–2.39, –0.37], *r* = 0.495) (for more details see **Table 3B**).

Frustration Correlated Negative With Dominance, Control and Self-Confidence

After the first (Block 2, T2) as well as the second (Block 2, T4) interaction task with the two robots a positive correlation between frustration and arousal was found (T2: [0.295, 0.789], *r* = 0. 578; T4: [0.140, 0. 842], *r* = 0.562). The correlations between frustration and the (respective) parameters dominance, control, self-confidence, and self-reported task performance are negative after both interaction tasks. The higher the frustration score, the lower the dominance score, the sense of control and self-report task performance (T2: dominance: [–0.685, –0.128], *r* = –0.459; control: [–0.779, –0.410], *r* = –0.601; self-confident: [–0.754, –0.322], *r* = –0.576; T4: dominance: [–0.690, –0.076], *r* = –0.445; control: [–0.842, –0.440], *r* = –0.673; self-confident: [–0.858, –0.530], *r* = –0.717). The subjects rated their task performance worse when frustration was high (T4: [–0.844, –0.228], *r* = –0.587) (for more details see also **Table 3B**).

Robot Rating: Robots Were Evaluated Different in Condition Groups

After each interaction task both the robot and the interaction were evaluated with the Post-Study System Usability Questionnaire (**Figure 3**: Block 2, T2 and T4). **Figure 5** shows the evaluation of each robot ("*Pepper*" and "*Panda*") independent of the interaction sequence.

Both robot systems were rated better in the NOFRUST-group than in the FRUST-group independent of the sequence of interaction task (**Figures 5A,B** and see also **Table 5**). In the NOFRUST-group the robots were evaluated very similarly except for the category "correction of errors." In the FRUST-group the robot "*Panda*" was rated worse than "*Pepper,*" except for the category "satisfaction" and "LED-feedback."

Participants in the NOFRUST-group described "*Pepper*" as more manageable and found it easier to correct its errors in both interaction tasks compared to the FRUST-group. In addition, the participants found "*Pepper*" easier to brief than in the FRUST-group (**Figure 6A**). Participants found the interaction task with "*Panda*" more productive and the robot easier to use in the NOFRUST-group than in the FRUST-group (**Figure 6B**). "*Panda*" was rated worse in more categories in the FRUST-group than in the NOFRUST-group and then "*Pepper*" in the FRUST-group (**Figures 6A,B** and see also **Table 5**).

There was no significant difference between the two robots in the FRUST-group on the indication of frustration and overall perception after the 1st interaction task (frustration: *MD* = –6.54 [–38.42, 25.34], *r* = 0.115; overall perception: *MD* = 0.47 [–0.67, 1.62], *r* = 0.224).

The robot "*Panda*" was rated more negatively than "*Pepper*" in the second interaction in the FRUST group in the following categories: easy to use (*MD* = 1.38 [0.06, 2.7], *r* = 0.506), pleasant use (*MD* = 1.14 [0.21, 2.06], *r* = 0.576), productivity (*MD* = 1.58 [0.28, 2.89], *r* = 0.576), fun (*MD* = 1.35 [0.42, 2.27], *r* = 0.655), overall perception (*MD* = 1.94 [0.89, 3.1], *r* = 0.726).

After the two interaction tasks, participants indicated which robot they preferred and why. "*Pepper*" was described as more human-like. It was attributed to be more trustworthy and "enabled more familiar interactions" with a more pleasant feeling. About "*Panda*" they stated that it was more functional, it was limited to the bare minimum of functionality, and the behavior was more expectable. More subjects preferred to interact with "*Pepper.*"

DISCUSSION

In this paper, results on the influence of frustration in a HRI study were presented and are discussed in the following section about recommendations for successful HRI.

Short Description of the Study Design

In the reported study, participants performed a task in collaboration with a robotic system and with a common goal in a handover scenario. The participants interacted with two different robot systems, one after the other. There were two condition groups, frustration (FRUST) and no frustration (NOFRUST). Frustration was successfully induced through technical errors.

Summary of the Results

In this section, the results are considered in relation to the hypotheses (see section "Hypotheses").

H1: Technical Errors By the Robot Lead to Perceived Frustration By the Participants

The results showed that frustration occurred in the FRUST-group in both interaction tasks (with both robots). The operationalization of frustration was successful, also seen in the questionnaires (NASA-TLX and EaCQ). In the videos, reactions were found mainly in the faces of the participants, especially laughter and smiling. This is also reflected in statistical differences in the condition groups.

Participants defined frustration remarkably similar. Frustration is mainly associated with "disappointment," especially with "expectations," and "not reaching a goal." These terms also correspond to the definition seen in several definitions in the literature in the introduction section (e.g., Freud, 1921; Russell, 1980; Amsel, 1992; Bortz and Doering, 2013).

Since facial expressions were very often shown in association with frustration, these might be good candidates to detect frustration in interaction situations. Usually specific facial expressions are expected, e.g., indicating frustration (Jost, 1941; Scherer, 2005; Hamm et al., 2011; Hazlett, 2013; Gao et al., 2014; Lerner et al., 2015). Therefore, a more careful way of detecting emotions is necessary, including the situational context. Detecting smiles and laughter by emotion detectors will not reflect the entire situation if the context of reoccurring failure is not considered.

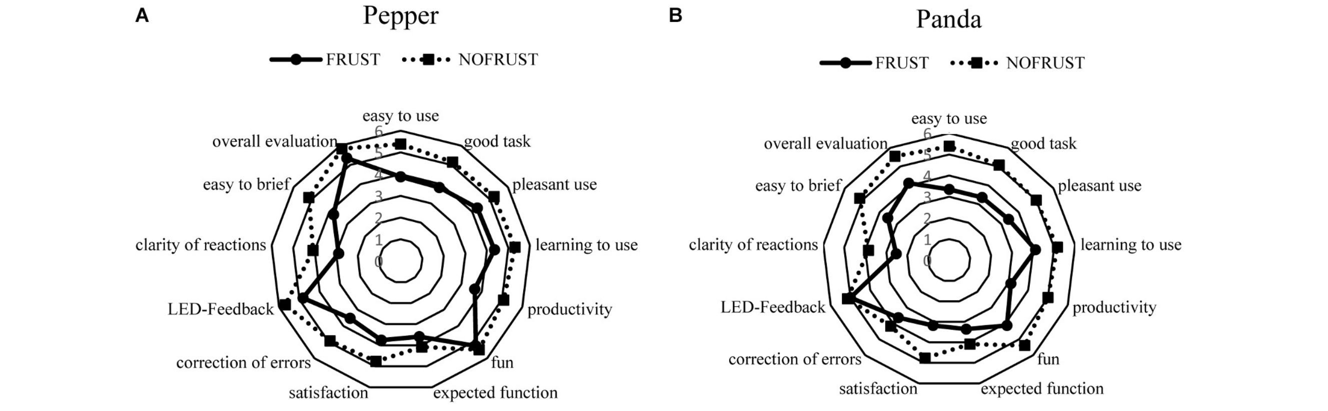


FIGURE 5 | Evaluation of the robot systems [“*Pepper*” **(A)** And “*Panda*” **(B)**] independent of the sequence of interaction task.

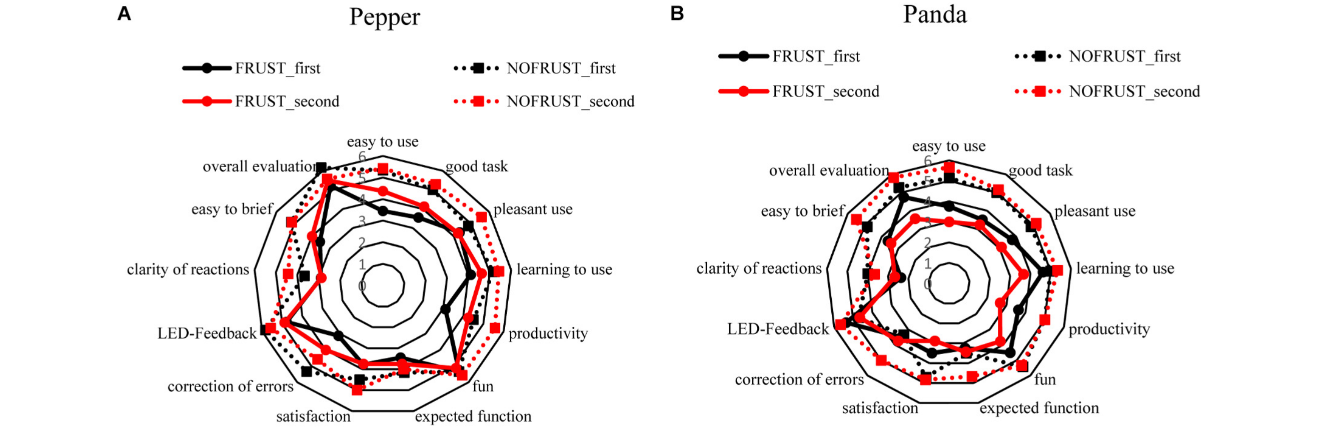


FIGURE 6 | Evaluation of the robot systems [“*Pepper*” **(A)** And “*Panda*” **(B)**] after the first and second interaction task.

H2: Frustration Leads to Decreased Dominance, Sense of Control, and Self-reported Performance

Frustration affects dominance, the sense of control and self-confidence in both interaction situations. Frustration has shown negative correlation with all three characteristics.

As shown in other studies, frustration has an influence on interaction factors. We find the sense of dominance and control in interaction particularly relevant, which is very important for the evaluation of system and interaction quality and thus for a good collaboration. To be able to assess the situation is important for joint task accomplishment and collaboration as was mentioned earlier.

H3: Frustration Leads to Lower Rating Regarding Acceptance of the Robot Systems

H4: The Interaction With the More Human-like Robot (Here “*Pepper*”) is Preferred, Among Other Aspects Due to the Human-like Appearance and Similarity to HHI

The robots have been evaluated differently in the condition groups, especially in the categories “easy to use,” “productivity,” and “easy to brief.” In the FRUST-group the robots were rated more negatively than in the NOFRUST-group. The robot “*Pepper*” was rated more positively on average than the robot arm “*Panda.*”

Frustration has an impact on the evaluation of interaction and robot systems. The experience of frustration seems to have a negative impact on the evaluation of the easy handling and the possibility to give good instructions. Thus, the task cannot be fulfilled as expected which in turn leads to disappointed expectations and frustration.

No significant difference was found between the frustration levels in the interaction tasks with the two robotic systems. Thus, the interaction task with the robot seemed to be independent in respect to the order of which robot is used first.

What Are Relevant Factors for a Successful Human–Robot Interaction?

The following will describe the aspects which were examined in this study, how the results can be interpreted, and what this could mean for future HRI research.

Appearance

The two systems in this study differed in their appearance to examine if the appearance has an influence on the interaction. Furthermore, movement could form different expectations and might have a strong influence on the interactive behavior of the participant and on the evaluation of the interaction task.

The participants had more confidence in familiar situations and found the interaction with "*Pepper*" more natural and less disconcerting, probably because the robot looked more human-like and thus evoked the expectations of a HHI. This led to a better assessment of the reactions and movements, which in turn can increase the sense of dominance and control. Riek et al. (2009) found a positive effect of anthropomorphism, they showed that people empathize more with robots which have a more human-like than a mechanical appearance (Onnasch and Roesler, 2019) and treated them differently (Malle et al., 2016). But the robot appearance preferences depend on the environmental context (e.g., home versus factory) (Chanseau, 2019) and the task. The relevant issue is how good the evoked expectations through the appearance can be fulfilled through the robot in the specific task.

The appearance of the robot "*Panda*" was rated more negatively in several categories by the FRUST-group than "*Pepper*" in the second interaction task. Frustration seems to have a negative influence on the evaluation of the interaction and the interaction partner. Participants indicated that they found it easier to interact with "*Pepper,*" the interaction was more fun, and they found the robot to be better in the overall interaction rating. When indicating which robot the subjects preferred to interact with, the subjects indicated "*Pepper*" more often.

Expectations and attributions based on the appearance of the robot and the environment as well as the task have a great influence on the interaction, albeit mostly subconsciously. Therefore, attention should be paid to the associations that appearance and previously known abilities of the robot have on human partners. But not just the first impression is important also the performance of the robot influences subjective perception of the robot (Salem et al., 2015).

The appearance and capabilities of the robot system should be adapted to the scope of the interaction, for example, in certain areas it should be limited to the most necessary aspects and be more functional. In addition, the speed of the system in the interaction is important, whereby human safety must be guaranteed, but the interaction should be pleasant and (possibly) natural.

The robots were rated differently in the condition groups. Thus, perceived frustration had a negative impact on the rating of the interaction and the interaction partner. More participants indicated that they preferred interacting with "*Pepper,*" mainly because of appearance and familiarity. Appearance seems to be an important aspect in HRI. Thus, the study indicated that humans like to work with familiar objects and that the appearance of robotic systems should be suited to the context of use and functionally appropriate. Of course, the expectation triggered by the appearance should not be ignored.

Behavioral Reaction to Robots

The occurrence of the specific facial expressions, smiling and laughing, during the interaction tasks especially in the FRUST-group with both robots was an interesting aspect in this study. This was also reflected in the correlation results between frustration and behavioral data from the videos.

The ability to recognize facial expressions as additional information about human experience in interaction is an interesting aspect for the design of a robot system. The facial expressions, such as laughter and smiles, can be misinterpreted by the robot system if facial expressions are not interpreted in the situational context.

Furthermore, the ability to interpret emotional reactions correctly could be a valuable information in social robotic systems that make use of concepts like joint attention and common goal representation. This information gives a hint if the assumed common goal and necessary actions are aligned by both partners. This provides means to correct the assumed instances to come back to a successful collaborative interaction which would release the possible frustrating experience of the partner.

The participants showed two different types of reactions to the robot's errors in interaction. The reactions were either directed toward the technical error or can be rated as attempts to correct the robot. Here, two types of errors can be differentiated, on the one hand traceable errors, which were more often treated with correction attempts, such as "hand-on-gestures" or color changes of the ball. On the other hand, non-traceable errors, whereupon only reactions, like facial expression, were shown. The error "gripper remains open" and "gripper remains closed" can be classified in the group of traceable errors, and the errors "accept ball and drop it in the human or robot area" are rather incomprehensible errors. Type of errors and intention of the robot influence anthrophomistic perception of the robot (Salem et al., 2015).

This shows the importance of research on cognitive modeling approaches that enable robots or intelligent systems to gain an understanding of the human partner (Kambhampati, 2019; Klaproth et al., 2020; Rußwinkel, 2020) in order to respond to the partner comprehensively. Furthermore, the robot needs to behave in a traceable fashion, so that the human partner is motivated to help even if errors occur. Just cases of pure "no comprehension" will be fatal for further interactions.

Thus, the study also showed that it is important to consider the behavior, especially the facial reactions of the human in the context of the interaction and that these are relevant for the course of the interaction. But without connecting the facial expression to the situation at hand, interpretations will remain difficult.

Dominance and Control

As shown in our previous study on frustration (Weidemann and Rußwinkel, 2019) the sense of dominance and control turn out to be important aspects in this context. The results revealed that frustration led to a reduction of sense of dominance and control.

The sense of dominance and control are important factors in an interaction and should be preserved for the human interaction partner (in the interaction). Negative emotions, such as discomfort, irritation, and frustration lead to the human partner to lose the sense of dominance and control which leads to a termination of the collaboration or at least to the negative evaluation of the interaction. Certainly, the acceptance of the robot system will decrease if the negative situation will not be solved.

Therefore, it is important to minimize negative emotions in the interaction. This can be achieved by for example fulfilling expectations, recognizing, and understanding human emotions and feelings, and showing the appropriate and desired feedback.

Feedback

An important aspect in the design of a good HRI is the feedback given by the robot to the human and also vice versa. For interpreting feedback reactions, it is important to understand if the partner has expected an event or agrees with the situation or decision of the partner.

For a good predictability, the robot should gather enough information about the human state and the human's action to interpret this information in the appropriate context. With human partners, a major part of communication relies on the facial expression. Humans give immediate feedback to the robot in form of reactions such as facial expressions (Lang et al., 2010). Finding ways of interpreting such immediate facial expressions under consideration of the current situation is a promising approach for designing better collaborative robotic systems.

In case the feedback from the robot to the human, i.e., would be adapted to the human's needs, would consider the situational context and would be accepted by the human, this would be considered as social feedback (Schneider, 2011). The feedback should serve the human being as support for the common fulfillment of the task, as well as representing the status and the next actions of the robot. This type of feedback can be realized through different channels, for example visual or haptic. LED feedback or other user interfaces are able to give immediate feedback.

The importance of interpreting and responding to facial expression was also demonstrated in this study. In addition, the use of LED-feedback helped in communication in fulfilling the common goal of the interaction task. Thus, the study showed that mutual feedback is important for pleasant HRI.

CONCLUSION

In this study we were able to successfully induce frustration in a collaborative HRI situation by errors made by robots that lead to frustration by the human interaction partner and a delay in achieving the common goal. This way, we were able to validate the results of Giuliani et al. (2015) and Abd et al. (2017). The setup and protocol used for the study could be used in further studies that investigate measurements of frustration or means of reducing frustration, e.g., by a careful design of feedback signals or other kind. As we have argued, such situations and the impact on the human partner has a serious influence on successful HRI. In addition, the study provided indications about aspects that should carefully be considered in designing a good interaction with a robot. These aspects are robot appearance and feedback reactions the robot should provide to diminish frustration response by the human partner.

If these aspects are included in future HRIs, robots are more likely to be accepted in human life and in the working world and thus can lead to an "integration" of robots.

Frustration was determined in this study using questionnaires and behavioral reactions. To better identify frustration, we also included psychophysiological data (electrocardiogram, electrodermal activity, and electromyogram) in the study. These can be recorded in parallel with the occurrence of frustration. Alongside our behavioral data, this data will be investigated, analyzed and discussed in more detail in future work. It will be beneficial to gain a deeper understanding what circumstances lead to frustration – may be even how the feeling of agency or sense of control can be supported in interaction situations.

This may provide the robot with additional data about the human's state during an interaction and allow it to recognize frustration or other emotions, and to respond appropriately. So they can help the human in his or her activities. Frustration can be minimized. How frustration can be minimized in a HRI should be investigated in future studies.

Another interesting question is whether and how the behavioral responses in the FRUST-group change over the interaction. Possibly, the changes of the frustration level could be determined by this data. This question should be investigated in another experimental design, also to measure frustration with different methods at multiple time points. Thus, parallel measurement to the occurrence of frustration with different methods and minimization of frustration should be investigated in future studies.

The fact that feedback between the interaction partners plays an important role was also made clear once again in this study. However, only the visual channel was considered in relation to the feedback by the robot (LED-Feedback). Which other or additional feedback channels are still suitable for HRI should be examined in further studies.

In general the study also provides evidence that it is of high relevance to consider emotional reactions in HRI which also provides information on the others expectations and motivations. This could be done by emotion recognition programs or by measuring arousal. But taken alone this will not help since the context of the situation the emotion changes is of high relevance for the interpretation. Emotional reactions therefore be considered as part of the communication that is taking part. Cognitive modeling could help to provide this kind of context as e.g., shown in neuro adaptive assistance systems or other approaches of human aware AI (Kambhampati, 2019).

In future studies the robot systems and interaction should be adapted according to the recommendations developed in this paper and be tested in interaction studies with similar tasks that take into account close interaction, feedback provided, evaluation of emotional reactions, behavioral data in non-functioning situations. Questions remain, how simple feedback reactions could lead to a better impression regarding sense of control. Or to find simpler methods to measure frustration and agency.

The main massage is, that more research is needed toward human aware robotic systems, modeling of the mental and cognitive state of the human partner for providing better anticipation skills, and to engage further into considering metrics for emotional reaction and interpretation.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee (EC) of the Institute for Psychology and Industrial Sciences (IPA), Technical University of Berlin. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AW and NR designed the experiment, drafted, edited, revised, and approved the manuscript. AW performed the experiments and analyzed the data. Both authors contributed to the article and approved the submitted version.

FUNDING

The research presented in this manuscript was primarily supported by the German Ministry of Education and Research.

REFERENCES

Abd, M. A., Gonzalez, I., Nojoumian, M., and Engeberg, E. D. (2017). "Trust, satisfaction and frustration measurements during human-robot interaction," in *Proceedings of the 30th Florida Conference on Recent Advances in RoboticsMay 11-12, 2107*, (Boca Raton, FL: Florida Atlantic University).

Amsel, A. (1992). *Frustration Theory: An Analysis of Dispositional Learning and Memory*. Cambridge: Cambridge University Press.

Bartneck, C., Croft, E., Kulic, D., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* 1, 71–81. doi: 10.1007/s12369-008-0001-3

Bartneck, C., Verbunt, M., Mubin, O., and Al Mahmud, A. (2007). "To kill a mockingbird robot," in *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, (New York, NY: ACM), 81–87.

Beer, J. M., Liles, K. R., Wu, X., and Pakala, S. (2017). "Chapter 15 - Affective human–robot interaction," in *Emotions and Affect in Human Factors and Human-Computer Interaction*, ed. M. Jeon (Cambridge, MA: Academic Press), 359–381.

Berger, C. R. (1994). "Power, dominance, and social interaction," in *Handbook of Interpersonal Communication*, 2nd Edn, eds M. L. Knapp and G. R. Miller (Thousand Oaks, CA: Sage), 450–507.

Bessière, K., Newhagen, J. E., Robinson, J. P., and Shneiderman, B. (2004). A model for computer frustration the role of instrumental and dispositional factors on incident, session, and post-session frustration and mood. *Comput. Hum. Behav.* 22, 941–961. doi: 10.1016/j.chb.2004.03.015

Bethel, C. L., and Murphy, R. R. (2010). Review of human studies methods in HRI and recommendations. *Int. J. Soc. Robot.* 2, 347–359. doi: 10.1007/s12369-010-0064-9

Bortz, J., and Doering, N. (2013). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler*. Berlin: Springer.

Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* 25, 49–59. doi: 10.1016/0005-7916(94)90063-9

Brave, S., and Nass, C. (2002). "Emotion in human-computer interaction," in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, eds A. Sears and J. A. Jacko (Mahwah, NJ: Erlbaum Associates Inc.), 81–96.

Broadbent, E., Tamagawa, R., Patience, A., Knock, B., Kerse, N., Day, K., et al. (2012). Attitudes towards health-care robots in a retirement village. *Austral. J. Ageing* 31, 115–120. doi: 10.1111/j.1741-6612.2011.00551.x

Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., and Shneiderman, B. (2004). Determining causes and severity of enduser frustration. *Int. J. Hum. Comput. Interact.* 17, 333–356. doi: 10.1207/s15327590ijhc1703_3

Chanseau, A. (2019). *How People's Perception on Degree of Control Influences Human-Robot Interaction*. doctoral thesis, University of Hertfordshire, Hatfield .

Chen, J. S., Gross, K., Stanton, M., and Amsel, A. (1981). Adjustment of weanling and adolescent rats to a reward condition requiring slow responding. *Dev. Psychobiol.* 14, 139–145. doi: 10.1002/dev.420140207

Ciardo, F., De Tommaso, D., Beyer, F., and Wykowska, A. (2018). *Reduced Sense of Agency in Human-Robot Interaction*. London: ICSR, 441–450.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. Hillsdale, NJ: Erlbaum.

Compagna, D., Weidemann, A., Marquardt, M., and Graf, P. (2016). Sociological and biological insights on how to prevent the reduction in cognitive activity that stems from robots assuming workloads in human–robot cooperation, societies, MDPI. *Open Access J.* 6, 1–11.

Cumming, G. (2014). Die neue Statistik: warum und wie. *Psychol. Wissenschaft* 25, 7–29. doi: 10.1177/0956797613504966

Dautenhahn, K. (2007). Methodology and themes of human-robot interaction: a growing research field. *Int. J. Adv. Robot. Syst.* 2007, 103–108. doi: 10.5772/5702

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M. L., Koay, K. L., and Werry, I. (2005). "What is a robot companion-friend, assistant or butler?, in 'Intelligent Robots and Systems, 2005.(IROS 2005)," in *Proceedings of the 2005 IEEE/RSJ International Conference*, (Piscataway, NJ: IEEE), 1192–1197.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340. doi: 10.2307/249008

Fiore, S. M., Wiltshire, T. J., Lobato, E. J., Jentsch, F. G., Huang, W. H., and Axelrod, B. (2013). Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemics behavior. *Front. Psychol.* 4:859. doi: 10.3389/fpsyg.2013.00859

Freud, S. (1921). "Group psychology and the analysis of the ego," in *Proceedings of The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XVIII (1920-1922): Beyond the Pleasure Principle, Group Psychology and Other Works*, (London: The Hogarth Press and the Institute of Psychoanalysis) 65–14.

Gao, H., Yuce, A., and Thiran, J.-P. (2014). "Detecting emotional stress from facial expressions for driving safety," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, (Piscataway, NJ: IEEE), 5961–5965.

Gignac, G. E., and Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Pers. Individ. Diff.* 102, 74–78. doi: 10.1016/j.paid.2016.06.069

Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., and Tscheligi, M. (2015). Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Front. Psychol.* 6:931. doi: 10.3389/fpsyg.2015.00931

Graesser, A. C., Chipman, P., Haynes, B. C., and Olney, A. (2005). AutoTutor: an intelligent tutoring system with mixedinitiative dialogue. *Educ. IEEE Trans.* 48, 612–618. doi: 10.1109/TE.2005.856149

Haggard, P., and Chambon, V. (2012). Sense of agency. *Curr. Biol.* 22, R390–R392. doi: 10.1016/j.cub.2012.02.040

Haggard, P., and Tsakiris, M. (2009). The experience of agency feelings, judgments, and responsibility. *Curr. Direct. Psychol. Sci.* 18, 242–246. doi: 10.1111/j.1467-8721.2009.01644.x

Hamm, J., Kohler, C. G., Gur, R. C., and Verma, R. (2011). Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders. *J. Neurosci. Methods* 200, 237–256. doi: 10.1016/j.jneumeth.2011.06.023

Hart, S. G., and Staveland, L. E. (1988). ""Development of NASA-TLX (task load index): results of empirical and theoretical research" (PDF)," in *Human Mental Workload. Advances in Psychology*, eds P. A. Hancock and N. Meshkati (Amsterdam: North Holland), 139–183.

Hazlett, R. L. (2013). *Measurement of User Frustration: A Biologic Approach, CHI '03 Extended Abstracts on Human Factors in Computing Systems, April 05-10, 2003*. Ft. Lauderdale, FL: Springer.

Heerink, M., Kröse, B., Evers, V., and Wielinga, B. (2007). "Observing conversational expressiveness of elderly users interacting with a robot and screen agent," in *Proceedings of the IEEE 10th International Conference on Rehabilitation Robotics. (ICORR 2007)*, Berlin.

Hirshfield, L. M., Hirshfield, S. H., Hincks, S., Russell, M., Ward, R., and Williams, T. (2011). "Trust in human-computer interactions as measured by frustration, surprise, and workload," in *FAC 2011, HCII 2011, LNAI 6780*, eds D. D. Schmorrow and C. M. Fidopiastis (Berlin: Springer), 507–516.

Honig, S., and Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: literature review and model development. *Front. Psychol.* 9:861. doi: 10.3389/fpsyg.2018.00861

Ihme, K., Dömeland, C., Freese, M., and Jipp, M. (2018). Frustration in the face of the driver: a simulator study on facial muscle activity during frustrated driving. *Interact. Stud.* 19, 487–498.

Ihme, K., Unni, A., Rieger, J., and Jipp, M. (2016). "Assessing driver frustration using functional near infrared spectroscopy (fNIRS)," in *Proceedings of the 1st International Conference on Neuroergonomics (6 - 7Oct)*, Paris.

Ihme, K., Zhang, M., and Jipp, M. (2017). "Automatische erkennung der frustration von autofahrern: ergebnisse und anwendungsmöglichkeiten," in *Proceedings of the Automatisierungssysteme, Assistenzsysteme und eingebettete Systeme für Transportmittel*, (Iowa City, IA: AAET).

Janke, W., and Debus, G. (2003). "EWL eigenschaftwörterliste," in *Diagnostische Verfahren zu Lebensqualität und Wohlbefinden*, eds J. Schumacher, A. Klaiberg, and E. Brähler (Göttigen: Hogrefe), 92–96.

Janke, W., Debus, G., Erdmann, G., and Hüppe, M. (1995). *Befindlichkeitsskalierung Anhand von Kategorien und Eigenschaften*. Würzburg: Institut für Psychologie I. Unveröff. Fragebogen.

Janke, W., Hüppe, M., Kallus, W., and Schmidt-Atzert, L. (1988). *Befindlichkeitsskalierung Anhand von Kategorien und Eigenschaftswörtern (BSKE-E)*. Würzburg: Institut für Psychologie I. Unveröff. Fragebogen.

Jerčić, P., Wen, W., Hagelbäck, J., and Sundstedt, V. (2018). The effect of emotions and social behavior on performance in a collaborative serious game between humans and autonomous robots. *Int. J. Soc. Robot.* 10, 115–129. doi: 10.1007/s12369-017-0437-4

Jost, H. (1941). Some physiological changes during frustration. *Child Dev.* 12:9.

Kambhampati, S. (2019). Challenges of human-aware AI-systems. *AI Magazine* 41, 3–17.

Kiesler, S., and Hinds, P. (2004). Introduction to the special issue on human-robot interaction. *Hum. Comput. Interact.* 19, 101–102. doi: 10.1109/TSMCC.2004.826271

Klaproth, O. W., Halbrügge, M., Krol, L. R., Vernaleken, C., Zander, T. O., and Russwinkel, N. (2020). A neuroadaptive cognitive model for dealing with uncertainty in tracing pilots'. *Cogn. State. Top. Cogn. Sci.* 12, 1012–1029. doi: 10.1111/tops.12515

Klein, J., Moon, Y., and Picard, R. W. (2002). This computer responds to user frustration: theory, design, and results. *Interact. Comput.* 14, 119–140.

Klug, M. (2012). *Bachelor Thesis Marius Klug: Emotionsbasierte Mensch-Roboter-Interaktion*. Bachelor's thesis, University of Tuebingen, Tübingen.

Kort, B., Reilly, R., and Picard, R. W. (2001). *An Affective Model of Interplay Between Emotions and Learning Reengineering Educational Pedagogy— Building a Learning Companion*. Washington, DC: IEEE Computer Society.

Krohne, H. W., Egloff, B., Kohlmann, C.-W., and Tausch, A. (1996). Untersuchungen mit einer deutschen version der «positive and negative affect schedule» (PANAS). *Diagnostica* 42, 139–156. doi: 10.1037/t49650-000

Lang, C., Wachsmuth, S., Wersing, H., and Hanheide, M. (2010). "Facial Expressions as Feedback Cue in Human-Robot interaction – a comparison between human and automatic recognition performances," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops*, San Francisco, CA, 79–85.

Lazar, J., Jones, A., Hackley, M., and Shneiderman, B. (2006). Severity and impact of computer user frustration: a comparison of student and workplace users. *Interact. Comput.* 18, 187–207. doi: 10.1016/j.intcom.2005.06.001

Lerner, J. S., Li, Y., Valdesolo, P., and Kassam, K. S. (2015). Emotion and decision making. *Psychology* 66:115043. doi: 10.1146/annurev-psych-010213-115043

Lewis, J. R. (1992). "Psychometric evaluation of the post-study system usability questionnaire: the PSSUQ," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, (Los Angeles, CA: SAGE Publications), 1259–1260.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *Int. J. Hum. Comput. Interact.* 14, 463–488.

Lopatovska, I., and Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Inform. Process. Manag.* 47, 575–592. doi: 10.1016/j.ipm.2010.09.001

Malle, B. F., Scheutz, M., Forlizzi, J., and Voiklis, J. (2016). "Which robot am I thinking about?: the impact of action and appearance on people's evaluations of a moral robot," in *Proceedings of The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, (Piscataway, NJ: IEEE Press), 125–132.

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

Nass, C., Fogg, B. J., and Moon, Y. (1996). Can computers be teammates? *Int. J. Hum. Comput. Stud.* 45, 669–678. doi: 10.1006/ijhc.1996.0073

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. C. (1995). Can computer personalities be human personalities? *Int. J. Hum. Comput. Stud.* 43, 223–239. doi: 10.1145/223355.223538

Ng, S. H., and Bradac, J. J. (1993). *Power in Language: Verbal Communication and Social Influence*. Thousand Oaks, CA: Sage.

Onnasch, L., Maier, X., and Jürgensohn, T. (2016). *Mensch-Roboter-Interaktion - Eine Taxonomie fuer alle Anwendungsfaelle. 1*. Auflage: Dortmund: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.

Onnasch, L., and Roesler, E. (2019). Anthropomorphizing robots: the effect of framing in human-robot collaboration. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet.* 63, 1311–1315. doi: 10.1177/1071181319631209

Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche* 13, 1–30.

Peterson, S., Reina, C., Waldman, D., and Becker, W. (2015). Using physiological methods to study emotions in organizations. *Res. Emot. Organ.* 11, 3–27. doi: 10.1108/S1746-979120150000011002

Picard, R. W. (1997). "Does HAL cry digital tears? Emotions and computers," in *Hal's Legacy: 2001's Computer as Dream and Reality*, ed. D. G. Stork (Cambridge, MA: The MIT Press), 279–303.

Powers, S. R., Rauh, C., Henning, R. A., Buck, R. W., and West, T. V. (2011). The effect of video feedback delay on frustration and emotion communication accuracy. *Comput. Hum. Behav.* 27, 1651–1657. doi: 10.1016/j.chb.2011.02.003

Rasmussen, J. (1982). Human errors - a taxonomy for describing human malfunction in industrial installations. *J. Occup. Accid.* 4, 311–333. doi: 10.1016/0376-6349(82)90041-4

Reeves, B., and Nass, C. (1998). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, 1st Edn. Stanford, CA: CSLI Publications.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., and Robinson, P. (2009). "How anthropomorphism affects empathy toward robots," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, (New York, NY: ACM), 245–246.

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Rußwinkel, N. (2020). "Antizipierende interaktiv lernende autonome Agenten," in *Mensch-Roboter-Kollaboration*, ed. H. J. Buxbaum (Wiesbaden: Springer).

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). *Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust*. Piscataway, NJ: IEEE.

Sauro, J., and Lewis, J. R. (2012). *Quantifying the User Experience: Practical Statistics for User Research*. Waltham, MA: Elsevier.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Soc. Sci. Inf.* 44, 695–729. doi: 10.1177/0539018405058216

Schneider, S. (2011). *Exploring Social Feedback in Human-Robot Interaction During Cognitive Stress; Masterarbeit im Fach Intelligente Systeme an der Technischen Fakultät Universität Bielefeld*. master's thesis, Technischen Fakultät Universität Bielefeld, Bielefeld.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends Cogn. Sci.* 10, 70–76. doi: 10.1016/j.tics.2005.12.009

Smarr, C. A., Mitzner, T. L., Beer, J. M., Prakash, A., Chen, T. L., Kemp, C. C., et al. (2014). Domestic robots for older adults: attitudes, preferences, and potential. *Int. J. Soc. Robot.* 6, 229–247. doi: 10.1007/s12369-013-0220-0

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: toward a unified view. *MIS Q.* 27, 425–478. doi: 10.2307/30036540

Vyzas, E., and Picard, R. W. (1999). "Offline and online recognition of emotion expression from physiological data," in *Proceedings of the Workshop on Emotio/Based Agent Architectures- at the Third International Conference on Autonomous Agents*, Seattle, WA, 135–142.

Wainer, J., Feil-Seifer, D. J., Shell, D. A., and Mataric, M. J. (2007). "Embodiment and human-robot interaction: a task-based perspective," in *Proceedings of the RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, (Piscataway, NJ: IEEE), 872–877.

Walters, M. L., Syrdal, D. S., Dautenhahn, K., Te Boekhorst, R., and Koay, K. L. (2008). Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Auton. Robots* 24, 159–178. doi: 10.1007/s10514-007-9058-3

Washington, G., and Adviser-Jones, R. P. (2011). *Understanding the Impact of User Frustration Intensities on Task Performance Using a Novel Adaptation of the OCC Theory of Emotions*. dissertation thesis, The George Washington University, Washington, DC.

Waterhouse, I. K., and Child, I. L. (1953). Frustration and the quality of performance. *J. Pers.* 21, 298–311.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54, 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Waytz, A., Cacioppo, J., and Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* 5, 219–232. doi: 10.1177/1745691610369336

Weidemann, A., and Rußwinkel, N. (2019). "Investigation of frustration," in *Proceedings of the Mensch und Computer 2019*, eds F. Alt, A. Bulling, and T. Döring (Piscataway, NJ: IEEE), 819–824.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychol. Rev.* 92, 548–573.

Whinghter, L. J., Cunningham, C., Wang, M., and Burnfield, J. L. (2008). The moderating role of goal orientation in the workload-frustration relationship. *J. Occup. Health Psychol.* 13, 283–291.

Woolf, B. P., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., and Picard, R. W. (2009). Affect-aware tutors: recognizing and responding to student affect. *Int. J. Learn. Technol.* 4, 129–164. doi: 10.1504/IJLT.2009.028804

Wulfert, E., Roland, B. D., Hartley, J., Wang, N., and Franco, C. (2005). Heart rate arousal and excitement in gambling: winners versus losers. *Psychol. Addict. Behav.* 19, 311–316. doi: 10.1037/0893-164X.19.3.311

Yang, E. (2016). Mitigating user frustration through adaptive feedback based on human-automation etiquette strategies. *Graduate Theses Dissert.* 2016:15843. doi: 10.31274/etd-180810-5470

Yuan, J., Ding, N., Liu, Y., and Yang, J. (2014). Unconscious emotion regulation Nonconscious reappraisal decreases emotion related physiological reactivity during frustration. *Cogn. Emot.* 29, 1042–1053. doi: 10.1080/02699931.2014.965663

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 09 April 2021
doi: 10.3389/frobt.2021.640444

Can Robots Earn Our Trust the Same Way Humans Do? A Systematic Exploration of Competence, Warmth, and Anthropomorphism as Determinants of Trust Development in HRI

Lara Christoforakos [1]*, *Alessio Gallucci* [1], *Tinatini Surmava-Große* [1], *Daniel Ullrich* [2] and *Sarah Diefenbach* [1]

[1] Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany, [2] Department of Computer Science, Ludwig-Maximilians-Universität München, Munich, Germany

OPEN ACCESS

Edited by:
Martina Mara,
Johannes Kepler University of
Linz, Austria

Reviewed by:
Soheil Keshmiri,
Advanced Telecommunications
Research Institute International
(ATR), Japan
Giulia Perugia,
Uppsala University, Sweden

***Correspondence:**
Lara Christoforakos
lara.christoforakos@psy.lmu.de

Specialty section:
This article was submitted to
Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

Received: 11 December 2020
Accepted: 05 March 2021
Published: 09 April 2021

Citation:
Christoforakos L, Gallucci A,
Surmava-Große T, Ullrich D and
Diefenbach S (2021) Can Robots Earn
Our Trust the Same Way Humans Do?
A Systematic Exploration of
Competence, Warmth, and
Anthropomorphism as Determinants
of Trust Development in HRI.
Front. Robot. AI 8:640444.
doi: 10.3389/frobt.2021.640444

Robots increasingly act as our social counterparts in domains such as healthcare and retail. For these human-robot interactions (HRI) to be effective, a question arises on whether we trust robots the same way we trust humans. We investigated whether the determinants competence and warmth, known to influence interpersonal trust development, influence trust development in HRI, and what role anthropomorphism plays in this interrelation. In two online studies with 2 × 2 between-subjects design, we investigated the role of robot competence (Study 1) and robot warmth (Study 2) in trust development in HRI. Each study explored the role of robot anthropomorphism in the respective interrelation. Videos showing an HRI were used for manipulations of robot competence (through varying gameplay competence) and robot anthropomorphism (through verbal and non-verbal design cues and the robot's presentation within the study introduction) in Study 1 ($n = 155$) as well as robot warmth (through varying compatibility of intentions with the human player) and robot anthropomorphism (same as Study 1) in Study 2 ($n = 157$). Results show a positive effect of robot competence (Study 1) and robot warmth (Study 2) on trust development in robots regarding anticipated trust and attributed trustworthiness. Subjective perceptions of competence (Study 1) and warmth (Study 2) mediated the interrelations in question. Considering applied manipulations, robot anthropomorphism neither moderated interrelations of robot competence and trust (Study 1) nor robot warmth and trust (Study 2). Considering subjective perceptions, perceived anthropomorphism moderated the effect of perceived competence (Study 1) and perceived warmth (Study 2) on trust on an attributional level. Overall results support the importance of robot competence and warmth for trust development in HRI and imply transferability regarding determinants of trust development in interpersonal interaction to HRI. Results indicate a possible role of perceived anthropomorphism in these interrelations and support a combined consideration of these variables in future studies. Insights deepen the understanding of key variables and their interaction in trust dynamics in HRI and suggest possibly relevant design factors to enable appropriate trust levels and a resulting desirable HRI. Methodological and conceptual limitations underline benefits of a rather robot-specific approach for future research.

Keywords: human-robot interaction, trust, trust development, trustworthiness, competence, warmth, anthropomorphism, social robots

INTRODUCTION

Besides social interaction with other humans, we are increasingly confronted with innovative, intelligent technologies as our social counterparts. Social robots, which are specifically designed to interact and communicate with humans (Bartneck and Forlizzi, 2004), represent a popular example of such. They become more and more present within our everyday lives, e.g., in the field of healthcare (e.g., Beasley, 2012), but also in retail and transportation, and support us in daily tasks, like shopping or ticket purchase. Oftentimes their interaction design does not even allow a clear distinction from human counterparts, e.g., when they appear in the form of chatbots. Therefore, increasingly interacting with technology as a social counterpart in domains we have been used to cooperating with humans in, a question arises on whether we trust robots the same way we trust humans. Apart from levels of trust, this question also pertains to determinants of trust development. It thus seems worthwhile to look into theoretical foundations of trust development in interpersonal interaction, especially since trust builds a basic precondition for effective HRI (Hancock et al., 2011; van Pinxteren et al., 2019), and research in different contexts revealed a particular skepticism of machines compared to humans in trustworthiness (Dietvorst et al., 2015) and related variables such as cooperation (Merritt and McGee, 2012; Ishowo-Oloko et al., 2019), particularly relevant in consequential fields of application, such as medicine and healthcare (Promberger and Baron, 2006; Ratanawongsa et al., 2016).

In line with the general approach of transferring theories and models of interpersonal interaction to human-computer interaction (HCI) and human-robot interaction (HRI) (e.g., Gockley et al., 2006; Aly and Tapus, 2016), single studies have explored this approach with regard to trust (de Visser et al., 2016; Kulms and Kopp, 2018). Yet, they have mostly focused on single determinants and barely applied systematic manipulations of the determinants in question.

In psychological literature, a prominent conception regarding determinants of trust development is that of competence and warmth (e.g., Mayer et al., 1995; Fiske et al., 2007). The perception of both competence, i.e., an individual's capability and skills, and warmth, i.e., an individual's good intentions toward another (e.g., Mayer et al., 1995; Fiske et al., 2007), appear to foster development of trust in a human counterpart. In the context of HRI, single study results imply an according importance of similar determinants of trust development. Namely, in their metanalysis, Hancock et al. (2011) found that robot-related performance-based factors (e.g., reliability, false alarm rate, failure rate) were associated with trust development in HRI. Moreover, considering HCI in general, Kulms and Kopp (2018) have found that competence and warmth of a computer are positively related to trust development in computers.

Comparing trust in HRI to interpersonal trust, another possibly relevant determinant is anthropomorphism, namely the act of attributing human characteristics, motivations, emotions, and intentions to non-human agents (Epley et al., 2007). If we trust robots as we trust humans, the degree of a robot's human-likeness might also affect our trust in robots. Especially, since robots are increasingly being designed in an anthropomorphic way, HRI research on this determinant is currently growing. Particularly, recent studies have suggested humanlike robot design to be a promising strategy in fostering trust (e.g., Kiesler et al., 2008; Hancock et al., 2011). However, anthropomorphism has not been investigated in combination with other possible determinants to further clarify its role in trust development within HRI.

In sum, the assumingly relevant determinants of trust development in HRI, namely competence, warmth, and anthropomorphism, including their interactions, have not been comprehensively considered and systematically manipulated in HRI research. The purpose of our study was to systematically explore the transferability of determinants of interpersonal trust development (here: competence and warmth), further considering anthropomorphism as a possible influencing factor and exploring its interaction with the determinants in question. Specifically, we explored whether robot competence and warmth influence trust development in robots and what role anthropomorphism plays in this interrelation.

Results in this respect could contribute to HRI research by delivering deeper insights into conceptual relationships and underlying psychological mechanisms of trust development in HRI, shedding light on central variables and their interaction as well as examining the transferability of well-founded knowledge on interpersonal trust to HRI. Moreover, understanding what makes humans trust robots could come with implications on a societal level. It could foster a more reflected interaction with robots by highlighting reasons we trust robots in tasks such as dealing with our personal data. On a more practical level, based on the systematic manipulations of assumed relevant determinants of trust development in HRI, our research could offer insights on key design elements, which influence trust in robots and could thus be crucial in achieving desired trust levels within a particular HRI.

In the following sections we outline psychological theories and study results on determinants of interpersonal trust development, followed by recent research on determinants of trust development in HRI, reflecting on the transferability of insights. Afterwards, we present two studies each focusing on a separate combination of possible determinants of trust development in HRI and the according results and discussion. This is followed by a general discussion, considering overall limitations and future research.

TRUST DEVELOPMENT IN INTERPERSONAL INTERACTION AND HRI

As a multidimensional phenomenon, various definitions of trust can be found in the literature (e.g., Barber, 1983; Rempel et al., 1985; Rousseau et al., 1998). For example, in the context of technology-related trust, trust has been defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004, p. 54). Trust thus forms a basis for dealing with risk and uncertainty (Deutsch, 1962; Mayer et al., 1995) and fosters cooperative behavior (Corritore et al., 2003; Balliet and Van Lange, 2013). Although trust generally evolves over time and is based on multiple interactions (Rempel et al., 1985), especially in first encounters or short-time interactions, single trustee attributes may be crucial for attributed trustworthiness (e.g., Mayer et al., 2003).

Determinants of Trust Development in Interpersonal Interaction

The broadly applied Stereotype Content Model (Fiske et al., 1999, 2002) suggests that individuals' judgment of others can be classified by the two universal dimensions of social cognition: competence and warmth. Whereas competence represents "traits that are related to perceived ability," warmth stands for "traits that are related to perceived intent" (Fiske et al., 2007, p. 77). The authors propose that these dimensions can predict individuals' affective and behavioral responses (Fiske et al., 2007; Cuddy et al., 2008), such as the extent to which a trustor trusts the trustee. Therefore, the higher the perception of competence or warmth, the more positive the judgment, i.e., the higher the trust in the trustee.

Another model supporting the importance of these dimensions in interpersonal trust development is the widely accepted model by Mayer et al. (1995), describing trustee attributes and behaviors, such as trustworthiness, and trustor attributes, such as trust propensity, as essential determinants of trust development. Focusing on the trustee, the authors propose a three-factor model describing antecedents of trustworthiness, including ability, benevolence, and integrity. Ability represents the "group of skills, competencies, and characteristics that enable a party to have influence within some specific domain" (Mayer et al., 1995, p. 717). Benevolence represents the extent to which the trustor believes the trustee to have good intentions toward the trustor and integrity is given, when the trustor perceives that the trustee follows principles accepted by the trustor (Mayer et al., 1995). The higher these determinants are perceived, the higher the trustworthiness attributed to the trustee.

Recent study results also support the importance of similar determinants for trust development and social cognition overall. van der Werff and Buckley (2017) investigated trust development in co-worker relationships to identify cues that foster trusting behaviors. Results showed that competence and benevolence of the trustee were positively related to disclosure and reliance (van der Werff and Buckley, 2017) as forms of trust behavior (Gillespie, 2003).

Despite slightly varying terms (e.g., ability and benevolence, Mayer et al., 1995; competence and morality, Phalet and Poppe, 1997; competence and warmth, Fiske et al., 2007), competence and warmth seem to be central dimensions of individuals' perception of others. Focusing on trust, perceiving the trustee as capable of achieving certain intended goals (competence) as well as adhering to the same intentions and interests as the trustor (warmth) can foster trust development in interpersonal relationships (Mayer et al., 1995; Fiske et al., 2002, 2007).

Transferability of Determinants of Trust Development in Interpersonal Interaction to HRI

A popular definition of trust in HRI describes trust as a "belief held by the trustor that the trustee will act in a manner that mitigates the trustor's risk in a situation, in which the trustor has put its outcomes at risk" (Wagner, 2009, p. 31). As research on trust development in HRI is relatively recent, theoretical models and studies on trust in interpersonal interaction as well as HCI can act as fundamental groundwork. Moreover, the "computers are social actors" paradigm (Nass and Moon, 2000) specifies that individuals apply social heuristics from human interactions in HCI, supporting the relevance of findings in interpersonal trust for trust in HRI. Furthermore, empirical studies show a strong correlation of trust in robots with trust in automation (Parasuraman et al., 2008; Chen et al., 2010), supporting the applicability of results regarding trust in this context to HRI (Hancock et al., 2011).

Accordingly, parallel to interpersonal trust, numerous studies have found a relevance of determinants related to robot competence for trust development in HRI. These include the robot's perceived competence based on its facial expressions (Calvo-Barajas et al., 2020), the robot's reputation in the sense of knowledge about its reliance (Bagheri and Jamieson, 2004), its previous performance (Chen et al., 2010, Lee and See, 2004), as well as its actual performance (Chen et al., 2010). Similarly, Robinette et al. (2017) found that poor robot performance was associated with a drop in self-reported trust of humans in robots, which was in turn correlated with their decision to use the robot for guidance (Robinette et al., 2017). Furthermore, in their metanalysis Hancock et al. (2011) showed that robot-related performance-based factors, such as reliability, false-alarm rate, and failure rate, predicted trust development in robots. Thus, perceiving the trustee (the robot) as competent, i.e., capable of achieving intended goals, seems essential for trust development in HRI as well.

While in HRI research warmth has not been particularly investigated as a potential determinant of trust development, assumptions can be derived from HCI literature. For example, Kulms and Kopp (2018) examined the transferability of interpersonal trust dynamics in the domain of intelligent computers, focusing on competence and warmth as possible determinants of trust in such. Competence was manipulated by means of competent (vs. incompetent) gameplay of the computer and warmth by means of unselfish (vs. selfish) game behavior of the computer. Results showed that competence and warmth were positively related to trust in computers, implying a relevance and certain transferability of trust determinants from interpersonal trust to trust in HCI.

To what degree humans actually treat technologies as social counterparts (Reeves and Nass, 1996) and apply social heuristics from human interactions (Keijsers and Bartneck, 2018) also depends on the availability of social cues, e.g., a user interface or car front looking like a smile. Thus, regarding the transferability of interpersonal trust dynamics to HRI, anthropomorphism of robots might be a relevant determinant. Accordingly, study results support a positive relationship between anthropomorphic design cues, e.g., humanlike appearance or voice of robots (Hancock et al., 2011; van Pinxteren et al., 2019) as well as agents, in general, and trust in such (e.g., Pak et al., 2012; de Visser et al., 2016, 2017). Furthermore, Kulms and Kopp (2019) explored the role of anthropomorphism and advice quality, a sort of robot competence, in trust within a cooperative human-agent setting. Results support a positive effect of anthropomorphism on self-reported trust, but also imply that competence might be essential for behavioral trust. Overall, anthropomorphism as a possible contributing factor to trust development in HRI has mainly been considered in single empirical studies in HRI research and in combination with competence in a first study on HCI (Kulms and Kopp, 2019). Such results, as well as the possibly essential role of anthropomorphism in the transferability of interpersonal trust dynamics to HRI, support a combined consideration of anthropomorphism with competence and warmth as trust determinants in HRI. Specifically, anthropomorphism may moderate the effect of competence and warmth on trust in HRI by enhancing applicability of interpersonal trust dynamics to HRI.

HYPOTHESES AND RESEARCH PARADIGM

Based on theoretical approaches and recent findings, as summarized in the preceding paragraphs, our research explored the effect of robot competence and robot warmth on trusting a robot. We assumed that both determinants will enhance trust, focusing on two facets of trust, namely, anticipated trust toward the robot and attributed trustworthiness to the robot. We further hypothesized that this relation is mediated by individual perceptions of robot competence, which is characterized as robot warmth. In addition, we assumed that robot anthropomorphism may play a moderating role and could further strengthen the effect of robot competence and robot warmth on trust. These general hypotheses were explored in two consecutive experimental studies, each manipulating one of the possible trust determinants (Study 1: robot competence, Study 2: robot warmth). Both studies further investigated the possible moderating role of robot anthropomorphism and used the same robot and general study paradigm, consisting of experimental manipulations through a video of a specific HRI.

STUDY 1

Methods

Experimental Manipulation

A 2 × 2 between-subjects-design with manipulated competence (high vs. low) and manipulated anthropomorphism (high vs. low) as independent variables was applied.

For each experimental condition, a different interaction between a service robot and a human player was presented on video. In all videos the protagonists (robot and human player) were playing a shell game. The human player covered a small object with one of three shells and mixed up the shells with rapid movements. Afterwards, the robot guessed under which shell the object was hidden. Within all conditions four playthroughs were presented, all together lasting 1 min on average.

The manipulation of robot competence focused on the skills of the robot (e.g., Mayer et al., 1995; Fiske et al., 2007) in the shell game. In the condition with high competence, the robot's judgement was correct three out of four times. In the condition with low competence, the robot's judgment was correct one out of four times. Complete failure or success was avoided to allow variance within the perception of competence. To counter further possible confounding effects, e.g., of perceived warmth, the robot gave very brief answers (i.e., "left," "right"). Finally, the total game score was illustrated on the robot's tablet after the game to support participants' notice.

Based on study results regarding explicit and implicit cues that can foster anthropomorphism (e.g., Eyssel et al., 2011; Salem et al., 2013; Waytz et al., 2014), robot anthropomorphism was manipulated explicitly through verbal (voice) and non-verbal (gestures) design cues as well as implicitly through naming the robot within the introduction given to the study. In the condition with high anthropomorphism, the robot named "Pepper" showed the shell in question with its hand and moved its head in the according direction. In the condition with low anthropomorphism, the robot did not have a name, nor did it show any gestures, or speak. Instead, its answers were presented on its tablet.

For the videos, the service robot Pepper by SoftBank Mobile Corp. (Pandey and Gelin, 2018) was used. According to the Wizard-of-Oz method (Fraser and Gilbert, 1991), the robot's speech and gestures were remote-controlled and triggered using the software Choreograph for Windows. Furthermore, for the robot's speech the German male voice programmed for Apple's Siri was applied. Premiere Pro, Adobe was used for overall editing. Thereby, the human player's movements, while mixing up the shells, were sped up by 50%. To avoid possible contrast effects (Bierhoff and Herner, 2002), the human counterpart in the shell game was blurred out. The four conditions are described in **Table 1**. In **Figure 1**, screenshots of the videos in all four conditions are presented.

Participants

One hundred and fifty five participants between eighteen to seventy-seven years (M = 33.50 years, SD = 15.00 years; 63.87% female, 34.84% male, 1.29% diverse) took part in the study. Participants were mainly recruited via University mailing-lists and social media platforms. As an incentive for their participation, two gift coupons of thirty Euros were raffled among all participants. Alternatively, students could register their participation for course credit. There were no preconditions for participation.

Procedure

The study was realized via online questionnaire, using Unipark (EFS Fall 2019) for programming. The study was announced as a study on HRI. Participants were informed about the average duration of the study and available incentives. After participants informed consent regarding data privacy terms according to the German General Data Protection Regulation (DGVO) was obtained, they were randomly assigned to one of four experimental conditions. In each condition participants watched the video of the above-described HRI and afterwards provided different judgements on the robot and additional measures as further specified below. All measures were assessed in German, using pre-tested translations if no validated versions were available.

TABLE 1 | Descriptions of experimental conditions in study 1.

Experimental conditions	Competence high	Competence low
Anthropomorphism high	Video of shell game with robot “Pepper,” who is right in three out of four trials, speaks with a humanlike voice and points out the shell in question.	Video of shell game with robot “Pepper,” who is right in one out of four trials, speaks with a humanlike voice and points out the shell in question.
Anthropomorphism low	Video of shell game with robot, who is right in three out of four trials, presenting its answers written on its tablet’s screen without voice or gestures.	Video of shell game with robot, who is right in one out of four trials, presenting its answers written on its tablet’s screen without voice or gestures.

Experimental condition competence high x anthropomorphism high, n = 37.
Experimental condition competence high x anthropomorphism low, n = 41.
Experimental condition competence low x anthropomorphism high, n = 33.
Experimental condition competence low x anthropomorphism low, n = 44.

Measures

Anticipated Trust

Anticipated trust toward the robot as one measure of trust in our study was measured by the five-item Faith subscale of the measure for human-computer trust by Madsen and Gregor (2000) (e.g., If I am not sure about a decision, I have faith that the system will provide the best solution). Items were assessed on a seven-point Likert-Scale (1 = “does not apply at all”; 7 = “applies fully”) and showed an internal consistency of $\alpha = 0.88$.

Attributed Trustworthiness

Attributed trustworthiness to the robot as the second measure of trust in our study was measured by a six-item scale of terms for assessing trustworthiness as a dimension of credibility of computer products by Fogg and Tseng (1999). The item “well-intentioned” was excluded to minimize confounding effects with robot warmth. The resulting five items (i.e., trustworthy, good, truthful, unbiased, honest) were assessed on a five-point Likert-Scale (1 = “does not apply at all”; 5 = “applies fully”) and showed an internal consistency of $\alpha = 0.79$.

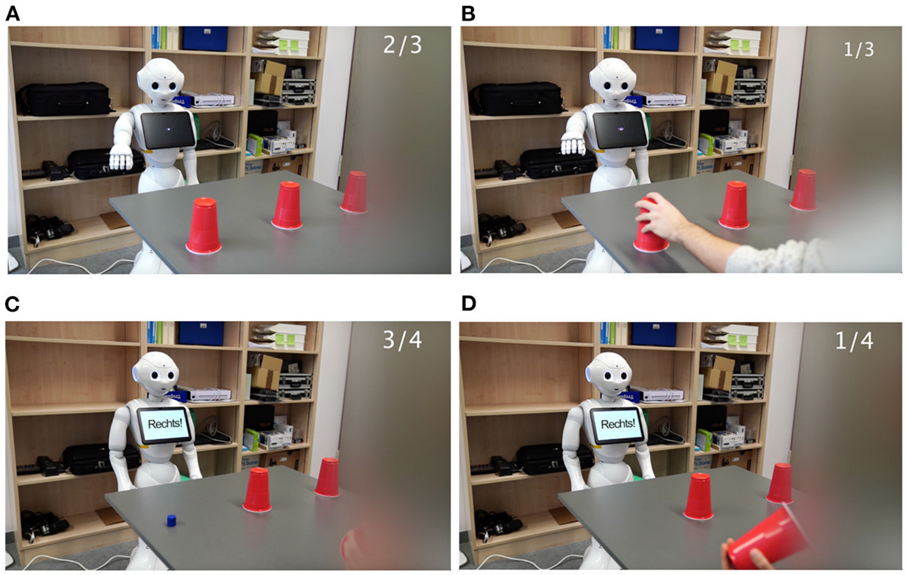


FIGURE 1 | Screenshots of the videos in Study 1, displaying HRI during a shell game in the conditions **(A)** anthropomorphism high x competence high, **(B)** anthropomorphism high x competence low, **(C)** anthropomorphism low, competence high, and **(D)** anthropomorphism low, competence low. Game scores are presented in the upper right corner of each screenshot.

Perceived Anthropomorphism

Participants' perceived anthropomorphism of the robot was measured by a single item (i.e., The robot made a humanlike impression), assessed on a five-point Likert Scale (1 = "does not apply at all"; 5 = "applies fully").

Perceived Competence

Participants' perceived competence of the robot was measured by means of the six-item Competence scale by Fiske et al. (2002), initially developed to assess stereotypes in interpersonal interaction. Items (i.e., competent, confident, capable, efficient, intelligent, skilful) were assessed on a seven-point Likert Scale (1 = "does not apply at all"; 7 = "applies fully") and showed an internal consistency of $\alpha = 0.84$.

Perceived Warmth

Participants' perceived warmth of the robot was measured by means of the six-item Warmth scale by Fiske et al. (2002), initially developed to assess stereotypes in interpersonal interaction. The item "trustworthy" was excluded to minimize confounding effects with attributed trustworthiness. The resulting five items (i.e., friendly, well-intentioned, warm, good-natured, sincere) were assessed on a seven-point Likert Scale (1 = "does not apply at all"; 7 = "applies fully") and showed an internal consistency of $\alpha = 0.93$.

Individual Tendency to Anthropomorphize

Participants' individual tendency to anthropomorphize was measured by means of the ten-item AQcurrent subscale of the Anthropomorphism Questionnaire by Neave et al. (2015). Items (e.g., I sometimes wonder if my computer deliberately runs more slowly after I shouted at it) were assessed on a seven-point Likert Scale (1 = "does not apply at all"; 7 = "applies fully") and showed an internal consistency of $\alpha = 0.86$.

Experience With Technology/Robots

Participants' experience with technology and robots were each measured by a self-constructed item (i.e., I generally consider my knowledge and skills in the field of technology/robots to be high). Items were assessed on a five-point Likert Scale (1 = "does not apply at all"; 5 = "applies fully").

Attitude Toward Robots

Participants' attitude toward robots was measured by means of the four-item Attitude Toward Robots subscale of the Robot Acceptance Questionnaire by Wu et al. (2014). Items (e.g., The robot would make life more interesting and stimulating in the future) were assessed on a five-point Likert Scale (1 = "does not apply at all"; 5 = "applies fully") and showed an internal consistency of $\alpha = 0.90$.

Demographic Measures

Participant's age was assessed by means of an open question. Gender was assessed through a single choice question with three answer options (i.e., male, female, diverse).

Hypotheses

Based on the above derived general hypotheses we specified the following for Study 1.

H1a: Individuals confronted with the robot with high competence (vs. low competence) will show higher anticipated trust.
H1b: Individuals confronted with the robot with high competence (vs. low competence) will attribute higher trustworthiness to the robot.
H2a: The effect of manipulated competence on anticipated trust is mediated through perceived competence of the robot.
H2b: The effect of manipulated competence on attributed trustworthiness is mediated through perceived competence of the robot.
H3a: The effect of manipulated competence on anticipated trust is strengthened by manipulated anthropomorphism.
H3b: The effect of manipulated competence on attributed trustworthiness is strengthened by manipulated anthropomorphism.

Results

Analyses were conducted with SPSS (IBM Statistics Version 26). For mediation and moderation analyses the Process Macro (Hayes and Preacher, 2013) was used.

Preliminary Analyses

Means, standard deviations, and Pearson correlations of the variables within the overall sample of Study 1 are illustrated in **Table 2**.

One-way ANOVAs showed no effect of the experimental conditions on age [$F_{(3,151)} = 0.69$, $p = 0.562$, $\eta^2 = 0.013$], individual tendency to anthropomorphize [$F_{(3,151)} = 0.39$, $p = 0.763$, $\eta^2 = 0.008$], experience with technology [$F_{(3,151)} = 0.50$, $p = 0.687$, $\eta^2 = 0.010$], experience with robots [$F_{(3,151)} = 1.01$, $p = 0.354$, $\eta^2 = 0.021$], or attitude toward robots [$F_{(3,151)} = 1.65$, $p = 0.180$, $\eta^2 = 0.032$]. The conducted Pearson's chi-squared test showed that experimental conditions did not differ significantly in gender distribution X^2 (6, $N = 155$) = 4.19, $p = 0.651$). Thus, there were no systematic differences regarding these variables to be further considered.

Furthermore, conducted one-way ANOVAs for manipulation checks showed that, as intended, manipulated competence had a significant effect on perceived competence [$F_{(1,153)} = 44.47$, $p < 0.001$, $\eta^2_p = 0.225$] as mean perceived competence was higher for conditions of high competence ($M = 4.18$, $SD = 1.26$) than low competence ($M = 2.90$, $SD = 1.12$). Additionally, according to our manipulation, manipulated anthropomorphism had a significant effect on perceived anthropomorphism [$F_{(1,153)} = 12.81$, $p < 0.001$, $\eta^2_p = 0.077$] as mean perceived anthropomorphism was higher for conditions of high anthropomorphism ($M = 2.56$, $SD = 1.16$) than low anthropomorphism ($M = 1.94$, $SD = 0.98$).

Hypotheses Testing

Two separate two-way ANOVAs were conducted to test the assumed effects of competence and anthropomorphism on anticipated trust (H1a, H3a) and attributed trustworthiness (H1b, H3b).

Regarding anticipated trust, the conducted two-way ANOVA showed a significant effect of manipulated competence [$F_{(3,151)}$

TABLE 2 | Means (*M*), standard deviations (*SD*), and Pearson correlations of relevant variables within the overall sample of study 1.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Age	33.5	15.00	–									
2. Anticipated trust	2.57	1.23	0.09	–								
3. Trustworthiness	2.97	0.86	−0.06	0.40**	–							
4. Perceived competence	3.55	1.34	−0.15	0.41**	0.69**	–						
5. Perceived anthropomorphism	2.22	1.11	−0.06	0.14	0.41**	0.25**	–					
6. Perceived warmth	3.45	1.53	−0.29**	0.14	0.46**	0.44**	0.39**	–				
7. Individual tendency to anthropomorphize	2.36	1.15	−0.27**	0.15	0.14	0.29**	0.11	0.27**	–			
8. Experience with technology	4.01	1.69	0.05	0.09	0.15	0.04	0.17*	0.17*	−0.07	-		
9. Experience with robots	2.61	1.68	0.08	0.16*	0.15	0.08	0.14	0.10	−0.02	0.73**	–	
10. Attitude toward robots	4.31	1.52	−0.08	0.16*	0.34**	0.27**	0.19*	0.31**	0.14	0.31**	0.25**	–

*Indicates $p < 0.05$.
**Indicates $p < 0.01$.

TABLE 3 | Mediated regression analysis testing the effect of manipulated competence on anticipated trust mediated by perceived competence within study 1.

					Model
Predictor	***B***	***SE***	***t***	***P***	***R*²**
Model 1: X on Y					0.14
Intercept	2.10	0.13	16.13	<0.001	
Manipulated competence	0.93	0.18	5.05	<0.001	
Model 2: X on M					0.23
Intercept	2.90	0.14	21.42	<0.001	
Manipulated competence	1.27	0.19	6.67	<0.001	
Model 3: X + M on Y					0.21
Intercept	1.30	0.25	5.19	<0.001	
Perceieved competence	0.28	0.08	3.70	<0.001	
Manipulated competence	0.58	0.20	2.87	0.005	

$= 25.64$, $p < 0.001$, $\eta^2_p = 0.145$] but not manipulated anthropomorphism [$F_{(3,151)} = 0.24$, $p = 0.602$, $\eta^2_p = 0.002$]. No interaction effect of manipulated competence and manipulated anthropomorphism on anticipated trust [$F_{(3,151)} = 0.681$, $p = 0.411$, $\eta^2_p = 0.004$] was found. Mean anticipated trust was higher for conditions of high competence ($M = 3.03$; $SD = 1.11$) compared to low competence ($M = 2.10$; $SD = 1.17$). Thus, H1a was supported. No moderation effect of manipulated anthropomorphism on the effect of manipulated competence on anticipated trust was found. Thus, H3a was not supported.

Regarding attributed trustworthiness, the conducted two-way ANOVA showed a significant effect of manipulated competence [$F_{(3,151)} = 17.01$, $p < 0.001$, $\eta^2_p = 0.102$] but not manipulated anthropomorphism [$F_{(3,151)} = 3.02$, $p = 0.085$, $\eta^2_p = 0.020$]. No interaction effect of manipulated competence and manipulated anthropomorphism on attributed trustworthiness [$F_{(3,151)} = 2.06$, $p = 0.153$, $\eta^2_p = 0.013$] was found. Mean attributed trustworthiness was higher for conditions of high competence ($M = 3.23$; $SD = 0.80$) compared to low competence ($M = 2.70$; $SD = 0.83$). Thus, H1a was supported. No moderation effect of manipulated anthropomorphism on the effect of manipulated competence on attributed trustworthiness was found. Thus, H3a was not supported.

The conducted mediated regression analysis showed a positive total effect of manipulated competence on anticipated trust ($B = 0.93$, $t = 5.05$, $p < 0.001$) and that perceived competence significantly mediated this interrelation with a positive indirect effect ($B = 0.35$). A bootstrap 95% CI around the indirect effect did not contain zero (0.14; 0.61). The direct effect of manipulated competence on anticipated trust remained significant ($B = 0.58$, $t = 2.87$, $p = 0.005$) after including the mediator variable, implying a partial mediation, and partially supporting H2a. A detailed overview of the mediated regression analysis is presented in **Table 3**.

The conducted mediated regression analysis showed a positive total effect of manipulated competence on attributed trustworthiness ($B = 0.53$, $t = 4.05$; $p < 0.001$) and that perceived competence significantly mediated this interrelation with a positive indirect effect ($B = 0.56$). A bootstrap 95% CI around the indirect effect did not contain zero (0.37; 0.78). The direct effect of manipulated competence on attributed trustworthiness became not significant ($B = -0.03$, $t = -0.28$, $p = 0.784$) after including the mediator variable, implying a complete mediation, and supporting H2b. A detailed overview of the mediated regression analysis is presented in **Table 4**.

Exploratory Analyses

Exploratory analyses were performed to detect possible interrelations between the studied constructs beyond our predefined hypotheses. Hence, we tested effects of manipulated competence on perceived anthropomorphism as well as effects of manipulated anthropomorphism on perceived competence. Two one-way ANOVAs showed no effect of manipulated competence on perceived anthropomorphism [$F_{(1,153)} = 0.55$, $p = 0.460$; $\eta^2_p = 0.004$] but a significant effect of manipulated anthropomorphism on perceived competence [$F_{(1,153)} = 4.28$, $p = 0.040$; $\eta^2_p = 0.027$]. Thereby, mean perceived competence was higher for conditions of high

TABLE 4 | Mediated regression analysis testing the effect of manipulated competence on attributed trustworthiness mediated by perceived competence within study 1.

					Model
Predictor	***B***	***SE***	***T***	***P***	***R*²**
Model 1: X on Y					0.10
Intercept	2.70	0.09	28.98	<0.001	
Manipulated competence	0.53	0.13	4.05	<0.001	
Model 2: X on M					0.23
Intercept	2.90	0.14	21.42	<0.001	
Manipulated competence	1.27	0.19	6.67	<0.001	
Model 3: X + M on Y					0.47
Intercept	1.41	0.14	9.90	<0.001	
Perceived competence	0.44	0.04	10.37	<0.001	
Manipulated competence	−0.03	0.11	−0.27	0.784	

TABLE 5 | Moderated regression analysis testing the effect of perceived competence on attributed trustworthiness moderated by perceived anthropomorphism within study 1.

					Model
Predictor	***B***	***SE***	***T***	***P***	***R*²**
Model					0.54
Intercept	0.65	0.28	2.33	0.021	
Perceived competence	0.53	0.08	6.96	<0.001	
Perceived anthropomorphism	0.42	0.12	3.55	<0.001	
Perceived competence * perceived anthropomorphism	−0.06	0.03	−2.00	0.047	

**stand for interaction.*

anthropomorphism ($M = 3.79$; $SD = 1.38$) compared to low anthropomorphism ($M = 3.35$; $SD = 1.29$).

Furthermore, we conducted moderation analyses in parallel to the assumed interaction effect between competence and anthropomorphism on trust (H3), however, this time considering the participants' subjective perceptions of robot competence and robot anthropomorphism instead of the experimental factors as predictors of trust. Regarding anticipated trust as one trust measure, only perceived competence showed as a significant predictor ($B = 0.38$, $t = 2.57$, $p = 0.011$), whereas perceived anthropomorphism ($B = 0.06$, $t = 0.25$, $p = 0.806$) and the interaction of perceived competence and perceived anthropomorphism ($B = -0.00$, $t = -0.07$, $p = 0.945$) did not. Perceived anthropomorphism therefore did not moderate the effect of perceived competence on anticipated trust. Regarding attributed trustworthiness as the other trust measure, perceived competence ($B = 0.53$, $t = 6.96$; $p < 0.001$), perceived anthropomorphism ($B = 0.42$, $t = 3.55$; $p < 0.001$), as well as the interaction of perceived competence and perceived anthropomorphism ($B = -0.06$, $t = -2.00$, $p = 0.047$), showed as significant predictors. Perceived anthropomorphism therefore moderated the effect of perceived competence on attributed trustworthiness. A detailed overview of the moderation analysis is presented in **Table 5**.

Discussion

The aim of Study 1 was to investigate the influence of robot competence on trust in HRI as well as the role of robot anthropomorphism in this interrelation. In this regard we manipulated robot competence and robot anthropomorphism in videos, in which a robot played a shell game with a human player. Based on the robot's behavior in this HRI, study participants provided two types of trust ratings, namely, anticipated trust toward the robot and attributed trustworthiness to the robot. In conformity with our hypotheses, manipulated competence had a significant positive effect on anticipated trust as well as attributed trustworthiness and both interrelations were (partially) mediated by perceived competence. Thus, according to our findings, robot competence appears to be a possible determinant of trust development in HRI, supporting the transferability of competence as a determinant of trust development in interpersonal interaction (e.g., Mayer et al., 1995; Fiske et al., 2007) to HRI. In addition, our results are compatible with previous HRI research (e.g., Hancock et al., 2011; Robinette et al., 2017), implying a positive effect of robot competence on trust in robots.

However, contrary to our hypotheses, manipulated anthropomorphism did not moderate the effect of manipulated competence on the trust ratings. This might be rooted in a rather restricted variance of anthropomorphism due to the manipulation based on the same robot, with the identical visual appearance in both conditions. Previous results that revealed an effect of anthropomorphic agent design have used stronger manipulations, e.g., comparing different types of agents, such as computers vs. avatars (e.g., de Visser et al., 2016). Yet, exploratory analyses revealed that the perception of the robot as anthropomorphic may still play a role, given that the individually perceived anthropomorphism (as well as perceived competence) predicted trust in the robot. In addition, the individually perceived anthropomorphism moderated the effect of perceived competence on attributed trustworthiness. In sum, this underlines the role of individual perception for the formation of psychological judgments such as trust and hints at a further consideration of robot anthropomorphism as a determinant of trust development in HRI, especially in combination with other known relevant determinants, such as competence. This finding can be considered in line with study results, showing that humans lose confidence in erring computers quicker than erring humans, highlighting the role of competence for trust in HCI as well as indicating a possible interaction of competence and anthropomorphism in this regard (Dietvorst et al., 2015). Similarly, previous results by de Visser et al. (2016) found that an increasing (feedback) uncertainty regarding a robot's performance during a task magnified the effect of agent anthropomorphism on trust resilience, i.e., a higher resistance to breakdowns in trust. The authors argue that "increasing anthropomorphism may create a protective resistance against future errors" (de Visser et al., 2016), indicating an interaction of robot competence and robot anthropomorphism. Our second study explored warmth as a further potential determinant of trust, again in combination with anthropomorphism.

STUDY 2

Methods

Experimental Manipulation

A 2 × 2 between-subjects-design with manipulated warmth (high vs. low) and manipulated anthropomorphism (high vs. low) as independent variables was applied.

For each experimental condition, a different interaction between a service robot and a human player was presented on video. In all videos the protagonists (a robot and two human players) were playing a shell game. This time, human player 1 covered a small object with one of three shells and mixed up the shells with rapid movements. Afterwards, human player 2 guessed under which shell the object was hidden. The robot was standing next to human player 2 and appearing to also observe the game. Within all conditions three playthroughs were presented, all together lasting 1 min on average. In the first playthrough human player 2 guesses wrongly without consulting the robot, in the two following playthroughs human player 2 expresses a guess and the robot additionally consults afterwards.

The manipulation of robot warmth focused on the intentions of the robot (Mayer et al., 1995; Fiske et al., 2007) regarding the shell game. In the condition with high warmth, the robot had the same intentions and interests as human player 2 (human player 2 winning at the shell game). This was expressed by the robot showing compassion after the first lost playthrough and offering help. In the following playthroughs the robot consults human player 2 correctly and cheers after each win. In the condition with low warmth, the robot had opposed intentions and interests to human player 2 (human player 2 losing at the shell game). This was expressed by the robot depreciating human player 2 after the first lost playthrough, yet offering help. Human player 2 accepts the robot's help but loses at the second playthrough because of the robot's misleading advice. The robot cheers gleefully. In the third playthrough the robot again advises human player 2 on the decision. Yet, human player 2 does not follow the robot's advice and decides correctly, which the robot gets miffed at. To counter further possible confounding effects, e.g., of perceived competence, the robot appeared to know the correct answer in both conditions, as a basis to help (warmth high) or mislead (warmth low) human player 2. In addition, human player 2 always expressed an assumption before consulting the robot. Robot anthropomorphism was again manipulated explicitly through verbal (voice) and non-verbal (gestures) design cues as well as implicitly through naming the robot within the introduction given to the study. In the condition anthropomorphism high, the robot named "Pepper" verbally expressed its advice. Furthermore, it turned its head in the direction of player 2 while speaking. In the condition with low anthropomorphism, the robot did not have a name, nor did it show any gestures or speak. Instead, its advice was presented on its tablet.

TABLE 6 | Descriptions of experimental conditions in study 2.

Experimental conditions	Warmth high	Warmth low
High anthropomorphism	Video of shell game with robot "Pepper" consulting player 2 according to the player's interest, speaking with a humanlike voice and turning its head toward player 2 while speaking.	Video of shell game with robot "Pepper" consulting player 2 against the player's interest, speaking with a humanlike voice and turning its head toward player 2 while speaking.
Low anthropomorphism	Video of shell game with robot consulting player 2 according to the player's interest, presenting its advice written on its tablet's screen without voice or gestures.	Video of shell game with robot consulting player 2 against the player's interest, presenting its advice written on its tablet's screen without voice or gestures.

Experimental condition warmth high x anthropomorphism high, n = 40; Experimental condition warmth high x anthropomorphism low, n = 37; Experimental condition warmth low x anthropomorphism high, n = 39; Experimental condition warmth low x anthropomorphism low, n = 41.

For the videos, the same service robot as in Study 1 was used and the same method, software, and voice were used for the robot's speech and gestures. Similarly, the same program as in Study 1 was used for overall editing. In Study 2, human player 1's movements were not sped up, to not make guessing correctly appear highly competent in itself and cause possible confounding effects. Again, the human counterparts in the shell game were blurred out. The four conditions are described in **Table 6**. In **Figure 2** screenshots of the videos in all four conditions are presented.

Participants

One hundred and fifty seven participants between eighteen to sixty-seven years (M = 34.53 years, SD = 13.88 years; 60.51% female, 39.49% male) took part in the study. Participant recruiting method and offered incentives were the same as in Study 1. Again, there were no preconditions for participation.

Procedure

The study procedure was the exact same as in Study 1, except one detail regarding the order of measures in the survey. Namely, perceived warmth was assessed before perceived competence.

Measures

The applied measures were the same as in Study 1. All scales showed satisfactory internal scale consistency (anticipated trust: $\alpha = 0.88$, attributed trustworthiness: $\alpha = 0.88$, perceived warmth: $\alpha = 0.94$, perceived competence: $\alpha = 0.84$, individual tendency to anthropomorphize: $\alpha = 0.83$, attitude toward robots: $\alpha = 0.91$).

Hypotheses

Based on the above derived general hypotheses we specified the following for Study 2.

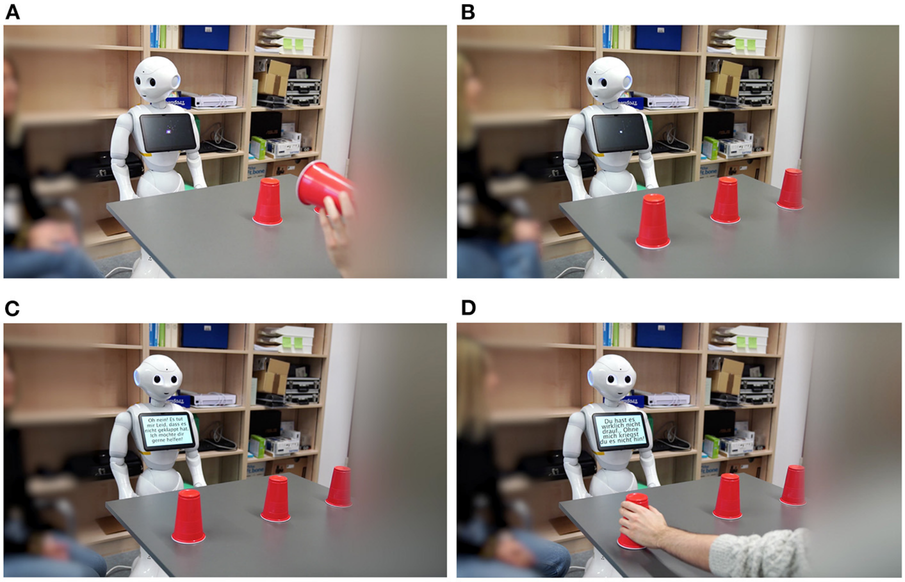

A
B
C
Oh nein! Es tut mir Leid, dass es nicht geklappt hat. Ich möchte dir gerne helfen!
D
Du hast es wirklich nicht drauf.. Ohne mich kriegst du es nicht hin!

FIGURE 2 | Screenshots of the videos in Study 2, displaying HRI during a shell game in the conditions **(A)** anthropomorphism high x warmth high, **(B)** anthropomorphism high x warmth low, **(C)** anthropomorphism low warmth high, and **(D)** anthropomorphism low warmth low.

H1a: Individuals confronted with the HRI with the robot with high warmth (vs. low warmth) will show higher anticipated trust.

H1b: Individuals confronted with the HRI with the robot with high warmth (vs. low warmth) will attribute higher trustworthiness to the robot.

H2a: The effect of manipulated warmth on anticipated trust is mediated through perceived warmth of the robot.

H2b: The effect of manipulated warmth on attributed trustworthiness is mediated through perceived warmth of the robot.

H3a: The effect of manipulated warmth on anticipated trust is strengthened by manipulated anthropomorphism.

H3b: The effect of manipulated warmth on attributed trustworthiness is strengthened by manipulated anthropomorphism.

Results

Analyses were conducted with SPSS (IBM Statistics Version 26). For mediation and moderation analyses the Process Macro (Hayes and Preacher, 2013) was used.

Preliminary Analyses

Means, standard deviations, and Pearson correlations of the variables within the overall sample of Study 2 are illustrated in **Table 7**.

One-way ANOVAs showed no effect of the experimental conditions on age [$F_{(3,153)} = 0.92$, $p = 0.431$, $\eta^2_p = 0.018$], individual tendency to anthropomorphize [$F_{(3,153)} = 1.71$, $p = 0.168$, $\eta^2_p = 0.032$], experience with robots [$F_{(3,153)} = 0.65$, $p = 0.568$, $\eta^2_p = 0.013$], experience with technology [$F_{(3,153)} = 0.70$, $p = 0.557$, $\eta^2_p = 0.013$], or attitude toward robots [$F_{(3,153)} = 1.18$, $p = 0.320$, $\eta^2_p = 0.023$]. The conducted Pearson's chi-squared test showed that experimental conditions did not differ significantly in gender distribution [$X^2{}_{(3,N=157)} = 1.79$, $p = 0.617$]. Thus, there were no systematic differences regarding these variables to be further considered.

Furthermore, conducted one-way ANOVAs for manipulation checks showed that, as intended, manipulated warmth had a significant effect on perceived warmth [$F_{(1,155)} = 62.63$, $p < 0.001$, $\eta^2_p = 0.288$] as mean perceived warmth was higher for conditions of high warmth ($M = 4.51$, $SD = 1.56$) than low warmth ($M = 2.64$, $SD = 1.40$). Additionally, according to our manipulation, manipulated anthropomorphism had a significant effect on perceived anthropomorphism [$F_{(1,155)} = 5.54$, $p = 0.020$, $\eta^2_p = 0.034$] as mean perceived anthropomorphism was higher for conditions of high anthropomorphism ($M = 2.66$, $SD = 1.26$) than low anthropomorphism ($M = 2.22$, $SD = 1.08$).

Hypotheses Testing

Two separate two-way ANOVAs were conducted to test the assumed effects of warmth and anthropomorphism on anticipated trust (H1a, H3a) and attributed trustworthiness (H1b, H3b).

Regarding anticipated trust, the conducted two-way ANOVA showed a significant effect of manipulated warmth [$F_{(3,153)} = 5.09$, $p = 0.026$, $\eta^2_p = 0.032$], but not manipulated anthropomorphism [$F_{(3,153)} = 0.30$, $p = 0.588$, $\eta^2_p = 0.002$]. No interaction effect of manipulated warmth and manipulated anthropomorphism on anticipated trust [$F_{(3,153)} = 2.67$, $p = 0.104$, $\eta^2_p = 0.017$] was found. Mean anticipated trust was higher for conditions of high warmth ($M = 3.40$; $SD = 1.46$) compared to low warmth ($M = 2.90$; $SD = 1.36$). Thus, H1a was supported.

TABLE 7 | Means (*M*), standard deviations (*SD*), and Pearson correlations of relevant variables within the overall sample of study 2.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Age	34.53	13.88	–									
2. Anticipated trust	3.14	1.43	0.16*	–								
3. Trustworthiness	2.78	1.07	0.09	0.45**	–							
4. Perceived warmth	3.55	1.75	0.12	0.33**	0.74**	–						
5. Perceived anthropomorphism	2.44	1.19	−0.05	0.14	0.27**	0.27**	–					
6. Perceived competence	4.08	1.38	−0.09	0.48**	0.49**	0.41**	0.32**	–				
7. Individual tendency to anthropomorphize	2.21	1.01	−0.10	0.17*	−0.02	0.02	0.21**	0.15	–			
8. Experience with technology	4.40	1.71	0.00	−0.03	0.02	0.04	−0.02	−0.12	0.03	–		
9. Experience with robots	2.82	1.67	0.06	0.15	0.11	0.09	0.02	−0.06	0.03	0.61**	–	
10. Attitude toward robots	4.10	1.60	0.17*	0.17*	0.08	0.11	0.05	0.05	0.04	0.26**	0.32**	–

*Indicates $p < 0.05$, **Indicates $p < 0.01$.

TABLE 8 | Mediated regression analysis testing the effect of manipulated warmth on anticipated trust mediated by perceived warmth in study 2.

					Model
Predictor	***B***	***SE***	***T***	***P***	***R*²**
Model 1: X on Y					0.03
Intercept	2.90	0.16	18.37	<0.001	
Manipulated warmth	0.50	0.23	2.22	0.28	
Model 2: X on M					0.29
Intercept	2.64	0.17	15.89	<0.001	
Manipulated warmth	1.87	0.24	7.91	<0.001	
Model 3: X + M on Y					0.11
Intercept	2.18	0.25	8.87	<0.001	
Perceieved warmth	0.27	0.07	3.72	<0.001	
Manipulated warmth	−0.01	0.26	−0.04	0.965	

TABLE 9 | Mediated regression analysis testing the effect of manipulated warmth on attributed trustworthiness mediated by perceived warmth in study 2.

					Model
Predictor	***B***	***SE***	***T***	***P***	***R*²**
Model 1: X on Y					0.30
Intercept	2.22	0.10	22.02	<0.001	
Manipulated warmth	1.16	0.14	8.05	<0.001	
Model 2: X on M					0.29
Intercept	2.64	0.17	15.89	<0.001	
Manipulated warmth	1.87	0.24	7.91	<0.001	
Model 3: X + M on Y					0.58
Intercept	1.20	0.13	9.50	<0.001	
Perceieved warmth	0.39	0.04	10.20	<0.001	
Manipulated warmth	0.43	0.13	3.30	0.001	

No moderation effect of manipulated anthropomorphism on the effect of manipulated warmth on anticipated trust was found. Thus, H3a was not supported.

Regarding attributed trustworthiness, the conducted two-way ANOVA showed a significant effect of manipulated warmth [$F_{(3,153)} = 63.83$, $p < 0.001$, $\eta^2_p = 0.294$] but not manipulated anthropomorphism [$F_{(3,153)} = 0.14$, $p = 0.708$, $\eta^2_p = 0.001$]. No interaction effect of manipulated warmth and manipulated anthropomorphism on attributed trustworthiness [$F_{(3,153)} = 0.06$, $p = 0.801$, $\eta^2_p < 0.001$] was found. Mean attributed trustworthiness was higher for conditions of high warmth ($M = 3.37$; $SD = 1.00$) compared to low warmth ($M = 2.22$; $SD = 0.79$). Thus, H1a was supported. No moderation effect of manipulated anthropomorphism on the effect of manipulated warmth on attributed trustworthiness was found. Thus, H3a was not supported.

The conducted mediated regression analysis showed a positive total effect of manipulated warmth on anticipated trust ($B = 0.50$, $t = 2.22$, $p = 0.028$) and that perceived warmth significantly mediated this interrelation with a positive indirect effect ($B = 0.51$). A bootstrap 95% CI around the indirect effect did not contain zero (0.22; 0.85). The direct effect of manipulated warmth on anticipated trust became not significant ($B = -0.01$, $t = -0.04$, $p = 0.965$) after including the mediator variable, implying a complete mediation, and supporting H2a. A detailed overview of the mediated regression analysis is presented in **Table 8**.

The conducted mediated regression analysis showed a positive total effect of manipulated warmth on attributed trustworthiness ($B = 1.16$, $t = 0.14$; $p < 0.001$) and that perceived warmth significantly mediated this interrelation with a positive indirect effect ($B = 0.72$). A bootstrap 95% CI around the indirect effect did not contain zero (0.12; 0.49). The direct effect of manipulated warmth on attributed trustworthiness remained significant ($B = 0.43$, $t = 3.30$, $p = 0.001$) after including the mediator variable, implying a partial mediation, and partially supporting H2b. A detailed overview of the mediated regression analysis is presented in **Table 9**.

Exploratory Analyses

Parallel to Study 1, exploratory analyses were performed to detect possible interrelations between the studied constructs beyond our predefined hypotheses. Hence, we tested effects of manipulated warmth on perceived anthropomorphism as well as effects of manipulated anthropomorphism on perceived warmth. Two

TABLE 10 | Moderated regression analysis testing the effect of perceived warmth on attributed trustworthiness moderated by perceived anthropomorphism within study 2.

					Model
Predictor	***B***	***SE***	***T***	***P***	***R*²**
Model					0.57
Intercept	1.55	0.28	5.55	<0.001	
Perceived warmth	0.28	0.08	3.52	<0.001	
Perceived anthropomorphism	−0.14	0.11	−1.30	0.196	
Perceived warmth * perceived anthropomorphism	0.06	0.03	2.17	0.032	

**stand for interaction.*

one-way ANOVAs showed no effect of manipulated warmth on perceived anthropomorphism [$F_{(1,155)} = 0.61$, $p = 0.435$; $\eta^2 = 0.004$] as well as no effect of manipulated anthropomorphism on perceived warmth [$F_{(1,155)} = 2.79$, $p = 0.097$; $\eta^2 = 0.018$].

Similar to Study 1, we conducted moderation analyses in parallel to the assumed interaction effect between robot warmth and robot anthropomorphism on trust (H3), however, this time considering the participants' subjective perceptions of robot warmth and robot anthropomorphism instead of the experimental factors as predictors of trust. Regarding anticipated trust as one trust measure, only perceived warmth showed as a significant predictor ($B = 0.36$, $t = 2.37$, $p = 0.019$), whereas perceived anthropomorphism ($B = 0.21$, $t = 0.97$, $p = 0.334$) and the interaction of perceived warmth and perceived anthropomorphism ($B = -0.04$, $t = -0.74$; $p = 0.460$) did not. Perceived anthropomorphism, therefore, did not moderate the effect of perceived warmth on anticipated trust. Regarding attributed trustworthiness as the other trust measure, perceived warmth ($B = 0.28$, $t = 3.52$, $p < 0.001$) as well as the interaction of perceived warmth and perceived anthropomorphism ($B = 0.06$, $t = 2.17$, $p = 0.032$) showed as significant predictors, whereas perceived anthropomorphism did not ($B = -0.14$, $t = -1.30$; $p = 0.196$). Perceived anthropomorphism, therefore, moderated the effect of perceived warmth on attributed trustworthiness. A detailed overview of the moderation analysis is presented in **Table 10**.

Discussion

The aim of Study 2 was to investigate the influence of robot warmth on trust in HRI as well as the role of robot anthropomorphism in this interrelation. In this regard, we manipulated robot warmth and robot anthropomorphism in videos, in which a robot consulted a human player in a shell game. In parallel to Study 1, based on the robot's behavior in this HRI, study participants provided two types of trust ratings, namely, attributed trustworthiness to the robot and anticipated trust toward the robot. In conformity with our hypotheses, manipulated warmth had a significant positive effect on anticipated trust as well as attributed trustworthiness and both interrelations were (partially) mediated by perceived warmth. Thus, according to our findings, robot warmth appears to be a possible determinant of trust development in HRI, supporting the transferability of warmth as a determinant of trust development in interpersonal interaction (e.g., Mayer et al., 1995; Fiske et al., 2007) to HRI. In addition, our results are compatible with previous HCI research (e.g., Kulms and Kopp, 2018), implying a positive effect of computer warmth on trust in computers.

Contrary to our hypotheses, manipulated anthropomorphism did not moderate the effect of manipulated warmth on the trust ratings. As elucidated in Study 1, a possible reason for this finding might be the restricted variance of anthropomorphism, due to its rather weak manipulation, based on the use of the same robot, with identical visual appearance in both conditions. Yet, exploratory analyses indicate that the perception of the robot as anthropomorphic may still play a role in this interrelation, when considering participants subjective perceptions of the determinants in questions. Namely, results showed that the individually perceived anthropomorphism moderated the effect of perceived warmth on attributed trustworthiness. These results indicate a further consideration of robot anthropomorphism, specifically its subjective perception, as a possibly relevant determinant of trust development in HRI, to be explored in combination with other known relevant determinants, such as warmth.

GENERAL DISCUSSION

The aim of our studies was to investigate whether the determinants competence and warmth, known to influence the development of interpersonal trust (e.g., Mayer et al., 1995; Fiske et al., 2007), influence trust development in HRI, and what role anthropomorphism plays in this interrelation. This was explored by two separate studies, one manipulating competence and anthropomorphism of a robot, and one manipulating warmth and anthropomorphism of a robot. Overall results imply a positive effect of robot competence (Study 1), as well as robot warmth (Study 2) on trust development in robots on an anticipatory as well as attributional level. These determinants thus seem relevant for trust development in HRI and support a transferability of essential trust dynamics from interpersonal interaction (Mayer et al., 1995; Fiske et al., 2007) to HRI.

Furthermore, considering the applied manipulations in both studies, anthropomorphic design cues in the robot neither influenced the interrelations of robot competence and trust (Study 1) nor robot warmth on trust (Study 2) on an anticipatory or attributional level. Yet, when considering participants' perception of the manipulated variables, an according effect was found; perceived anthropomorphism appeared to further influence the positive effect of perceived competence on attributed trustworthiness in Study 1 and perceived warmth on attributed trustworthiness in Study 2.

Our present results, then, contribute to research on trust development in HRI by highlighting the relevance of robot competence and robot warmth. Such results shed further light on the transferability of determinants of trust development from interpersonal interaction to HRI. Therefore, our research somewhat paves the way to understanding the complex network of factors in trust development within HRI. On a practical level, our results demonstrate how small differences in design within one single robot can come with significant differences in perceptions of the essential variables: robot competence, warmth, and anthropomorphism. Furthermore, our results offer first insights on design cues, which influence trust in robots and can thus be adjusted to foster appropriate levels of trust in HRI. Accordingly, the demonstration of high performance in a robot, e.g., by completing a task, as well as presenting the robot to have the same intentions as the user, can foster trust development. Furthermore, a perception of human likeness in a robot, e.g., based on a humanlike design, should be considered, as it might influence positive effects of perceived competence and perceived warmth of a robot on trust on an attributional level.

However, literature increasingly underlines consequences of overtrust in robotic systems. Robinette et al. (2017), for example, found that participants followed a robot's lead during an emergency even when it had performed incorrectly in previous demonstrations as well as when they were aware that the robot was acting wrongly. From an ethical perspective, it appears necessary to not only focus on design to foster trust in HRI but rather facilitate appropriate levels of trust. Although a detailed discussion in this regard would go beyond the scope of this paper, methods to foster appropriate levels of trust (e.g., Ullrich et al., 2021) should be considered in combination with the present research.

LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

Some methodological limitations within our studies, as well as more general limitations of the present research paradigm, need to be considered. First, regarding our applied manipulations within both studies, a central methodological limitation is the use of videos due to the online character of the studies. Thus, participants did not experience real HRI. Additionally, the short-time demonstrations of HRI might not have formed an appropriate basis to observe a possible development of trust in the robot. Furthermore, the robot we used for our manipulations was a commercial one. Thus, we cannot exclude a possible influence of previous experiences and resulting subjective impressions regarding the robot-related variables of interest. Regarding our applied measures, a methodological limitation is the use of self-reported trust measures. In future studies actual trust behavior should be assessed to foster external validity of results.

On a conceptual level, we must reflect on the general limitations of investigating the psychological dynamics behind HRI by means of experimental studies. While the experimental manipulation of single (presumably relevant) variables, generally, provides high internal validity, one can question whether this reductionist approach is the most sensible to detect relevant influencing factors in a complex domain such as trust development in HRI. As also demonstrated in the present study, operationalizing a sensitive construct as trust development in HRI, as well as possible determinants in experimental online studies, is a rather difficult task and typically connected to many possible confounding effects. Such could be the choice of robot as well as previous experience with robots in general (e.g., Hancock et al., 2011). Additionally, the task the robot is confronted with, specifically its type and complexity, could further affect trust in the robot (e.g., Hancock et al., 2011). Furthermore, humans' intraindividual dispositions could play a role. Accordingly, many studies support an interrelation of the Big Five personality traits (John et al., 1991), conscientiousness, agreeableness, extraversion, and trust in robots (e.g., Haring et al., 2013; Rossi et al., 2018). Although our intended manipulations were successful in both studies, the systematic manipulation of the assumed determinants of trust development under study turned out rather challenging. As exploratory results in Study 1 suggest, our manipulation of robot anthropomorphism might have also had an influence on perceived competence of the robot. While this finding might hint at the rather complex interrelation of the determinants in question, in sum, we cannot be sure whether our manipulations actually captured what is at the heart of people's mental models of robots and the question of trust or distrust. In this sense, one could even question to what extent the utilization of models of interpersonal interaction is useful to explore what determines trust in robots.

Therefore, in addition to experimental studies built on models of interpersonal trust, a change of perspective to robots "as an own species" may form another source of valuable insights (see also Ullrich et al., 2020). In alignment with previous research on specifically robotic qualities that does not try to parallel but rather highlights robot's differences to humans in psychological variables (e.g., a robot's endless patience as a "superpower," Welge and Hassenzahl, 2016; Dörrenbächer et al., 2020), future research could consider trust models that are unique to HRI. Such an alternative research approach could facilitate a more straightforward result interpretation and shed light on HRI-specific interrelations, which might have to date been overlooked, as they have not been discussed in comparable domains such as interpersonal interaction and thus need first-time exploration.

CONCLUSION

Although research agrees on the importance of trust for effective HRI (e.g., Freedy et al., 2007; Hancock et al., 2011; van Pinxteren et al., 2019), robot-related determinants of trust development in HRI have barely been considered or systematically explored. Comparing trust in HRI to interpersonal trust, our results imply a certain transferability of competence and warmth as central determinants of trust development in interpersonal interaction (e.g., Mayer et al., 1995; Fiske et al., 2007) to HRI, and hint at a possible role of the subjective perception of anthropomorphism in this regard.

While our research offers a valuable contribution to insights on trust dynamics in HRI, it also comes with methodological and conceptual limitations. Future studies could further attempt to optimize systematic manipulations of the found, relevant determinants of trust development in HRI and investigate such in a common study by additionally ensuring real life interaction with a robot, also measuring trust behavior. On a conceptual level, a question arises of whether experimental studies and the general utilization of models from interpersonal interaction represent a suitable approach to explore a complex domain such as trust development in HRI. It might thus be promising for future research to surpass existing models of trust, e.g., from interpersonal interaction, and focus on innovative approaches that are unique to HRI and highlight robot-specific interrelations.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors on https://data.ub.uni-muenchen.de/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements.

AUTHOR CONTRIBUTIONS

LC, DU, and SD conceived and planned the study. LC, AG, TS-G, and DU carried out the study and performed data analyses. All authors discussed the results and contributed to the manuscript.

FUNDING

This research was funded by the German Federal Ministry for Education and Research (BMBF), Project GINA (FKZ: 16SV8097).

ACKNOWLEDGMENTS

We thank Cedric Quintes for the programming of the robot used for the manipulations within the studies.

REFERENCES

Aly, A., and Tapus, A. (2016). Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction. *Autonomous Robots* 40, 193–209. doi: 10.1007/s10514-015-9444-1

Bagheri, N., and Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. In: *2004 IEEE International Conference on Systems, Man and Cybernetics Vol. 1. (IEEE Cat. No. 04CH37583)* IEEE. p. 212–7. doi: 10.1109/ICSMC.2004.1398299

Balliet, D., and Van Lange, P. A. (2013). Trust, conflict, and cooperation: a meta-analysis. *Psychol. Bull.* 139, 1090–1112. doi: 10.1037/a0030939

Barber, B. (1983). *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.

Bartneck, C., and Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. In *RO-MAN 2004.* In: *13th IEEE International Workshop on Robot and Human Interactive Communication* (IEEE Catalog No. 04TH8759). 591–594. IEEE. doi: 10.1109/ROMAN.2004.1374827

Beasley, R. A. (2012). Medical robots: current systems and research directions. *J. Robot.* 2012:401613. doi: 10.1155/2012/401613

Bierhoff, H. W., and Herner, M. J. (2002). *Begriffswörterbuch Sozialpsychologie*. Kohlhammer.

Calvo-Barajas, N., Perugia, G., and Castellano, G. (2020). "The effects of robot's facial expressions on children's first impressions of trustworthiness," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 165–171. doi: 10.1109/RO-MAN47096.2020.9223456

Chen, J. Y., Barnes, M. J., and Harper-Sciarini, M. (2010). Supervisory control of multiple robots: human-performance issues and user-interface design. *IEEE Transact. Syst. Man Cybernet. Part C* 41, 435–454. doi: 10.1109/TSMCC.2010.2056682

Corritore, C. L., Kracher, B., and Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. *Int. J. Human-Computer Stud.* 58, 737–758. doi: 10.1016/S1071-5819(03)00041-7

Cuddy, A. J., Fiske, S. T., and Glick, P. (2008). Warmth and competence as universal dimensions of social perception: the stereotype content model and the BIAS map. *Adv. Exp. Soc. Psychol.* 40, 61–149. doi: 10.1016/S0065-2601(07)00002-0

de Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R., et al. (2017). A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Human Fact.* 59, 116–133. doi: 10.1177/0018720816687205

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., et al. (2016). Almost human: anthropomorphism increases trust resilience in cognitive agents. *J. Exp. Psychol.* 22, 331–349. doi: 10.1037/xap0000092

Deutsch, M. (1962). "Cooperation and trust: Some theoretical notes," in *Nebraska Symposium on Motivation*, eds M. R. Jones (Lincoln, NB: University of Nebraska), 275–315.

Dietvorst, B., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol.* 144, 114–126. doi: 10.1037/xge0000033

Dörrenbächer, J., Löffler, D., and Hassenzahl, M. (2020). "Becoming a robot-overcoming anthropomorphism with techno-mimesis," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI), 1–12.

Epley, N., Waytz, A., and Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychol Rev.* 114:864. doi: 10.1037/0033-295X.114.4.864

Eyssel, F., Kuchenbrandt, D., and Bobinger, S. (2011). "Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism," in *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne), 61–68. doi: 10.1145/1957656.1957673

Fiske, S. T., Cuddy, A. J., and Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 77–83. doi: 10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J. Personal. Soc. Psychol.* 82:878. doi: 10.1037/0022-3514.82.6.878

Fiske, S. T., Xu, J., Cuddy, A. C., and Glick, P. (1999). (Dis) respecting versus (dis) liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *J. Soc. Issues* 55, 473–489. doi: 10.1111/0022-4537.00128

Fogg, B. J., and Tseng, H. (1999). "The elements of computer credibility," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 80–87. doi: 10.1145/302979.303001

Fraser, N. M., and Gilbert, G. N. (1991). Simulating speech systems. *Comput. Speech Language* 5, 81–99. doi: 10.1016/0885-2308(91)90019-M

Freedy, A., DeVisser, E., Weltman, G., and Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In: *2007 International Symposium on Collaborative Technologies and Systems* (IEEE), 106–114. doi: 10.1109/CTS.2007.4621745

Gillespie, N. (2003). *Measuring trust in work relationships: The Behavioural Trust Inventory*. Paper presented at the annual meeting of the Academy of Management. Seattle, WA.

Gockley, R., Simmons, R., and Forlizzi, J. (2006). Modeling affect in socially interactive robots. In: *ROMAN 2006-The 15th IEEE International*

Symposium on Robot and Human Interactive Communication (IEEE), 558–563. doi: 10.1109/ROMAN.2006.314448

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Fact.* 53, 517–527. doi: 10.1177/0018720811417254

Haring, K. S., Matsumoto, Y., and Watanabe, K. (2013). "How do people perceive and trust a lifelike robot," in *Proceedings of the World Congress on Engineering and Computer Science* (Vol. 1).

Hayes, A. F., and Preacher, K. J. (2013). *Conditional Process Modeling: Using Structural Equation Modeling to Examine Contingent Causal Processes.*

Ishowo-Oloko, F., Bonnefon, J. F., Soroye, Z., Crandall, J., Rahwan, I., and Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nat. Machine Intelligence* 1, 517–521. doi: 10.1038/s42256-019-0113-5

John, O. P., Donahue, E. M., and Kentle, R. L. (1991). Big five inventory. *J. Personal. Soc. Psychol.*

Keijsers, M., and Bartneck, C. (2018). "Mindless robots get bullied," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: ACM/IEEE), 205–214. doi: 10.1145/3171221.3171266

Kiesler, S., Powers, A., Fussell, S. R., and Torrey, C. (2008). Anthropomorphic interactions with a robot and robot–like agent. *Soc. Cognit.* 26, 169–181. doi: 10.1521/soco.2008.26.2.169

Kulms, P., and Kopp, S. (2018). A social cognition perspective on human–computer trust: the effect of perceived warmth and competence on trust in decision-making with computers. *Front. Digital Humanit.* 5:14. doi: 10.3389/fdigh.2018.00014

Kulms, P., and Kopp, S. (2019). "More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation," in *Proceedings of Mensch und Computer* (Hamburg), 31–42. doi: 10.1145/3340764.3340793

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Fact.* 46, 50–80. doi: 10.1518/hfes.46.1.50_30392

Madsen, M., and Gregor, S. (2000). "Measuring human-computer trust," in *Proceedings of Eleventh Australasian Conference on Information Systems*, 53 (Brisbane: QUT), 6–8.

Mayer, J. D., Salovey, P., Caruso, D. R., and Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion* 3, 97–105. doi: 10.1037/1528-3542.3.1.97

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 709–734. doi: 10.5465/amr.1995.9508080335

Merritt, T., and McGee, K. (2012). "Protecting artificial team-mates: more seems like less," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, TX), 2793–2802. doi: 10.1145/2207676.2208680

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Neave, N., Jackson, R., Saxton, T., and Hönekopp, J. (2015). The influence of anthropomorphic tendencies on human hoarding behaviours. *Personal. Individual Diff.* 72, 214–219. doi: 10.1016/j.paid.2014.08.041

Pak, R., Fink, N., Price, M., Bass, B., and Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics* 55, 1059–1072. doi: 10.1080/00140139.2012.691554

Pandey, A. K., and Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: the first machine of its kind. *IEEE Robot. Automation Magazine* 25, 40–48. doi: 10.1109/MRA.2018.2833157

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *J. Cognit. Eng. Decision Making* 2, 140–160. doi: 10.1518/155534308X284417

Phalet, K., and Poppe, E. (1997). Competence and morality dimensions of national and ethnic stereotypes: a study in six eastern-European countries. *Eur. J. Soc. Psychol.* 27, 703–723. doi: 10.1002/(SICI)1099-0992(199711/12)27:6<703::AID-EJSP841>3.0.CO;2-K

Promberger, M., and Baron, J. (2006). Do patients trust computers? *J. Behav. Decision Making* 19, 455–468. doi: 10.1002/bdm.542

Ratanawongsa, N., Barton, J. L., Lyles, C. R., Wu, M., Yelin, E. H., Martinez, D., et al. (2016). Association between clinician computer use and communication with patients in safety-net clinics. *JAMA Internal Med.* 176, 125–128. doi: 10.1001/jamainternmed.2015.6186

Reeves, B., and Nass, C. I. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University press.

Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *J. Personal. Soc. Psychol.* 49:95. doi: 10.1037/0022-3514.49.1.95

Robinette, P., Howard, A. M., and Wagner, A. R. (2017). Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transact. Human-Machine Syst.* 47, 425–436. doi: 10.1109/THMS.2017.2648849

Rossi, S., Santangelo, G., Staffa, M., Varrasi, S., Conti, D., and Di Nuovo, A. (2018). "Psychometric evaluation supported by a social robot: personality factors and technology acceptance," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (IEEE), 802–807. doi: 10.1109/ROMAN.2018.8525838

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Acad. Manage. Rev.* 23, 393–404. doi: 10.5465/amr.1998.926617

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. (2013). To err is human (-like): effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robot.* 5, 313–323. doi: 10.1007/s12369-013-0196-9

Ullrich, D., Butz, A., and Diefenbach, S. (2020). The eternal robot: anchoring effects in humans' mental models of robots and their self. *Front. Robot. AI*. 7:546724. doi: 10.3389/frobt.2020.546724

Ullrich, D., Butz, A., and Diefenbach, S. (2021). The development of overtrust: an empirical simulation and psychological analysis in the context of human-robot interaction. *Front. Robot*. doi: 10.3389/frobt.2021.554578

van der Werff, L., and Buckley, F. (2017). Getting to know you: a longitudinal examination of trust cues and trust development during socialization. *J. Manage. c*43, 742–770. doi: 10.1177/0149206314543475

van Pinxteren, M. M., Wetzels, R. W., Rüger, J., Pluymaekers, M., and Wetzels, M. (2019). Trust in humanoid robots: implications for services marketing. *J. Services Market*. 33, 507–518. doi: 10.1108/JSM-01-2018-0045

Wagner, A. R. (2009). *The Role of Trust and Relationships in Human-Robot Social Interaction*. Georgia Institute of Technology.

Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* 52, 113–117. doi: 10.1016/j.jesp.2014.01.005

Welge, J., and Hassenzahl, M. (2016). "Better than human: about the psychological superpowers of robots," in *International Conference on Social Robotics* (Cham: Springer), 993–1002. doi: 10.1007/978-3-319-47437-3_97

Wu, Y. H., Wrobel, J., Cornuet, M., Kerherv,é, H., Damnée, S., and Rigaud, A. S. (2014). Acceptance of an assistive robot in older adults: a mixed-method study of human–robot interaction over a 1-month period in the Living Lab setting. *Clin. Intervent. Aging* 9:801. doi: 10.2147/CIA.S56435

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 12 April 2021
doi: 10.3389/fpsyg.2021.592711

More Than a Feeling—Interrelation of Trust Layers in Human-Robot Interaction and the Role of User Dispositions and State Anxiety

Linda Miller†, Johannes Kraus*†, Franziska Babel and Martin Baumann*

Department Human Factors, Institute of Psychology and Education, Ulm University, Ulm, Germany

With service robots becoming more ubiquitous in social life, interaction design needs to adapt to novice users and the associated uncertainty in the first encounter with this technology in new emerging environments. Trust in robots is an essential psychological prerequisite to achieve safe and convenient cooperation between users and robots. This research focuses on psychological processes in which user dispositions and states affect trust in robots, which in turn is expected to impact the behavior and reactions in the interaction with robotic systems. In a laboratory experiment, the influence of propensity to trust in automation and negative attitudes toward robots on state anxiety, trust, and comfort distance toward a robot were explored. Participants were approached by a humanoid domestic robot two times and indicated their comfort distance and trust. The results favor the differentiation and interdependence of dispositional, initial, and dynamic learned trust layers. A mediation from the propensity to trust to initial learned trust by state anxiety provides an insight into the psychological processes through which personality traits might affect interindividual outcomes in human-robot interaction (HRI). The findings underline the meaningfulness of user characteristics as predictors for the initial approach to robots and the importance of considering users' individual learning history regarding technology and robots in particular.

Keywords: trust in robots, human-robot interaction, trust layers, user dispositions, affect, state anxiety, comfort distance, trust in automation

OPEN ACCESS

Edited by:
Myounghoon Jeon,
Virginia Tech, United States

Reviewed by:
EunJeong Cheon,
Aalborg University, Denmark
Shabnam FakhrHosseini,
Massachusetts Institute of Technology, United States

***Correspondence:**
Linda Miller
linda.miller@uni-ulm.de
Johannes Kraus
johannes.kraus@uni-ulm.de

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Emotion Science, a section of the journal Frontiers in Psychology

Received: *07 August 2020*
Accepted: *27 January 2021*
Published: *12 April 2021*

Citation:
Miller L, Kraus J, Babel F and Baumann M (2021) More Than a Feeling—Interrelation of Trust Layers in Human-Robot Interaction and the Role of User Dispositions and State Anxiety. Front. Psychol. 12:592711. doi: 10.3389/fpsyg.2021.592711

INTRODUCTION

Once utopian, robots are increasingly finding their way into public and private settings to assist humans in everyday tasks. Thereby, service robots offer numerous potentials for improvements in many fields, for example, by supporting disabled people to live more independently (e.g., Robinson et al., 2014). In these upcoming environments, robots represent a rather new and unfamiliar technology that most people have no specific knowledge or personal experience with. As many of these application areas for robots are characterized by increased complexity, dynamic, and interaction with untrained novice users, the interaction design needs to account for more flexibility and adaptability to both changing surroundings and users. Regarding the adaptability to users, it is a specifically important endeavor to reduce uncertainties and negative psychological consequences to facilitate an appropriate and repeated interaction with robots.

Based on interaction norms between humans, people treat robots as social partners in many respects. Thus robots are expected to behave in a socially acceptable manner and comply with social rules to some extent (e.g., Computers Are Social Actors paradigm, Nass et al., 2000; Nass and Moon, 2000; Rosenthal-von der Pütten et al., 2014). Thereby, amongst others, user characteristics (e.g., personality, Walters et al., 2005) were found to influence the individual reaction to robots. For example, such individual differences for the preferred proximity are discussed in Leichtmann and Nitsch (2020). At this point in human-robot interaction (HRI) design, psychological mechanisms need consideration to achieve positive interaction outcomes.

A multitude of research emphasized the importance of *trust* in the initial encounter with automated technologies (Lee and See, 2004; Stokes et al., 2010; Hoff and Bashir, 2015). Building on that, this research examines the role of this psychological variable that has also been thoroughly discussed and investigated in regard to the interaction with robots (e.g., Hancock et al., 2011; Schaefer, 2013; Salem et al., 2015), conceived in this context as advanced, complex automated technical systems. The research field of trust in automated systems is, amongst others, rooted in the observation that people do not use automation appropriately (Lee and See, 2004). Inappropriate use can be reflected in too much trust (overtrust), leading to misuse of a system on the one hand and too little trust (distrust), leading to disuse of a system on the other hand (Parasuraman and Riley, 1997; Lee and See, 2004). Thereby, to achieve an optimal, efficient, and safe interaction instead of an unconditional maximization of trust, designers might aim to achieve a calibrated level, which corresponds to a system's actual capabilities (calibrated trust; Muir, 1987). A calibrated level of trust has been related to a balanced usage of and reliance to innovative autonomous technology, thus facilitating a successful long-term relationship.

A good deal of research on trust in automation has focused on aviation and automated driving systems. While many of the findings in these and related areas might be readily transferred to HRI, this research seeks to validate and extend previous findings on the role of trust in automation to the interaction with robotic systems in domestic surroundings. Thereby, several specificities of domestic HRI have to be considered. First, domestic robots enter the user's personal space—not only in the sense of operating in private homes but also in a spatial and proxemic way. Second, robots can move around more flexible and might manipulate objects. Third, the prototypical user is not a trained professional. Fourth, domain-specific individual preferences, attitudes, and emotions are discussed to play a role in processes for evaluating and adopting robots. Essentially, negative robot attitudes and fear of robots are commonly discussed as potential influencing factors for the adoption of robots (Nomura et al., 2008; Syrdal et al., 2009; Złotowski et al., 2017). Still, the relationship of these factors with trust in robots has, up to now, only scarcely been investigated. Taken together, these particularities of HRI have to be kept in mind when comparing and transferring findings from other domains to the interaction with (domestic) robots.

In the presented study, the role of user dispositions and individual experiences of anxiety in the face of an unfamiliar robot on trust formation and proximity preferences has been investigated. This research focus is based on the overall assumption that general user dispositions affect the experiences throughout an individual's learning history with technology. This, on the one hand, leads to the formation of more specific technology-related personality traits and attitudes. On the other hand, the manifestation of an individual's dispositions in system-specific attitudes and behavior is expected to be subject to fluctuations and shaped by the affective state in a situation. In the presented laboratory experiment, users encountered a domestic service robot and indicated their state anxiety, trust, and comfort distance toward the robot. Based on recent findings on the role of individual levels of anxiety in the familiarization process with unfamiliar automated driving systems (e.g., Kraus et al., 2020a), this research is the first of its kind to extend the investigation of such a relationship to the domain of service robots. On this basis, suggestions for the design of initial interactions with robots in domestic environments are derived.

THEORETICAL BACKGROUND

Trust in Automation and Robots

Like in interpersonal relationships, trust is a fundamental requirement for successful human-machine interaction guiding decisions in unknown and risky situations (e.g., Lee and See, 2004; Hoff and Bashir, 2015). This is reflected in the definition of trust in automation as "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004, p. 51). On a conceptual level, trust is assumed to influence the behavior in regard to automation as part of a dynamic feedback loop, while the automation's attributes and actions also affect the level of trust (Lee and See, 2004). To establish an enhanced understanding of the complex psychological processes in which trust is formed, calibrated, and related to reliance decisions, a differentiation between several trust concepts seems worthwhile.

In their review on trust in automation, Hoff and Bashir (2015) conceptually distinguish different trust layers. Based on Kraus' (2020) *Three Stages of Trust* framework, an integration and extension of Lee and See's (2004) and Hoff and Bashir's (2015) models, three trust layers can be distinguished: the propensity to trust in automation, initial learned trust, and dynamic learned trust. First, the propensity to trust in automation (dispositional trust in Hoff and Bashir, 2015) refers to an automation-specific form of dispositional trust. The latter was defined in the interpersonal domain as "a diffuse expectation of others' trustworthiness [. . .] based on early trust-related experiences" (Merritt and Ilgen, 2008, p. 195). Building on this, the propensity to trust refers to a general context- and situation-independent personal predisposition to trust in automated technology (e.g., Hoff and Bashir, 2015). In the Three Stages of Trust framework, it is proposed that this trust layer is established from the combined influences of users' dispositions (e.g., demographics, culture, personality, and general technology attitudes) and the individual learning history with technology. Accordingly, users with a comparatively higher level of the propensity to trust in automation are more likely to be more trusting in the evaluation and interaction with unfamiliar automated systems, for example, robots (Kraus, 2020).

In contrast to the propensity to trust, *learned trust* comprises trust in a specific system. In learned trust, available information about a given system's trustworthiness is used to assess the system's trustworthiness in a situation of uncertainty (Kraus, 2020). This trustworthiness expectation is considerably informed by available diagnostic information—so-called trust cues—which were described as "in some way observable or given pieces of evidence a trustor might use to draw inferences about a trustee's trustworthiness in a specific situation" (Thielmann and Hilbig, 2015, p. 21). Kraus (2020)—based on Lee and Moray (1992) and Thielmann and Hilbig (2015)—proposes that the available information during trust formation can be differentiated into five categories along with the included trust cues: reputation-, purpose-, process-, performance-, and appearance-related. In the trust calibration process, the acquisition of new information affects the level of learned trust to the extent to which it derivates from the current trustworthiness expectation.

Within learned trust, one can further distinguish between *initial learned trust* based on information and existing knowledge prior to the interaction with a system and *dynamic learned trust*, which refers to trust adaptions during the actual interaction with a given system. It follows that learned trust is subject to change over time and is updated before and during the interaction by accumulated information and observations, for example, on perceived system performance (Merritt and Ilgen, 2008; Kraus et al., 2019b). It is further assumed that this process of learning to trust follows to a considerable extent the mechanisms of attitude formation and change (see Maio et al., 2018, exemplarily).

In the process of trust formation and calibration, the influences of many variables have been investigated in different technological domains. This is nicely summarized in several meta-analyses and reviews (see, e.g., Lee and See, 2004; Hancock et al., 2011; Hoff and Bashir, 2015; Schaefer et al., 2016). The different variables fall into the categories: person-related (e.g., personality, expertise, demographics), system-related (e.g., reliability, functionality, design), and situation-related (e.g., workload, affect). In the early days of research on trust in automation, the dynamic trust development in process control simulation micro-worlds was a central focus (e.g., Lee and Moray, 1992, 1994; Muir and Moray, 1996). More recently, trust processes were also investigated in the domains of automated driving (e.g., Hergeth et al., 2016), information technology (e.g., McKnight et al., 2002), and robots (e.g., Hancock et al., 2011).

This research's central focus is investigating the interrelation between the propensity to trust, initial learned trust before the interaction with a robot, and dynamic learned trust, developing and dynamically adapting during the interaction with a robot. The two forms of learned trust are thereby assumed to emerge in an attitude-formation process and then be calibrated along with a comparison of expectations with a robot's behavior. A detailed theoretical discussion of the psychological processes of formation and calibration of trust in automation is provided in the Three Stages of Trust framework (Kraus et al., 2019b; Kraus, 2020). A central assumption of this framework is that interindividual differences in trust are to some degree based on personality differences and the technology-related learning history of users, which affect the feelings toward and evaluation of a specific technological system (e.g., a robot).

Investigated Relationships

Following Kraus's (2020) framework, this research takes an integrative approach in the investigation of the interrelation of different trust layers (**Figure 1**). It is proposed that person characteristics (e.g., user dispositions and states) influence learned trust, which in turn builds a basis for behavior in HRI. In line with other trust models (e.g., Lee and See, 2004), the investigated model is rooted in the *Theory of Planned Behavior* (Fishbein and Ajzen, 1975; Ajzen, 1991), which assumes that behavior is determined by a cascade of beliefs, attitudes, and intentions. It is expected that characteristics of the situation (e.g., physical and legal attributes), the robot (e.g., appearance, performance, communication capabilities), and the task and interaction itself (e.g., goal, type) will moderate this process. Based on this, the hypotheses of this research are derived from theory and empirical findings in more detail below.

Trust as a Function of User Dispositions and States

This research builds on the general differentiation between cross-situational *traits*, which are comparatively stable person characteristics (e.g., personality), and short-term *states* (e.g., affect). States are situation-specific and reflect a person's adaption to given circumstances (e.g., social and physical situation, physiological and cognitive processes; Hamaker et al., 2007). Building on the assumption that traits and states contribute to interindividual variation in trust and behavior (e.g., Buss, 1989; Kraus, 2020), this research focuses on user characteristics as antecedents of trust. In line with this, trust in automation was found to be essentially influenced by both user traits and states (Merritt and Ilgen, 2008; Kraus et al., 2020a,b).

Above this, the inclusion of affect as a potential antecedent of trust in automation can be established on the basis of the *affect-as-information* model (e.g., Schwarz and Clore, 1988). This model proposes that people use their current affective state as an information basis for judgments about an object under consideration. Accordingly, various research has shown the impact of emotional and affective states on attention, perception, judgments, attitudinal responses, and behaviors in human interaction (e.g., Forgas and George, 2001; Brief and Weiss, 2002; Dunn and Schweitzer, 2005; Forgas and East, 2008). In conclusion, robot-related psychological outcomes such as trust are expected to be attenuated by users' affective states in a state-congruent direction (Brave and Nass, 2007). Taken together, it is proposed that users' state anxiety before and during an interaction affects their learned trust in a robot. This mechanism is expected to differ between users based on their dispositional propensity to trust in automation and attitude toward robots in general.

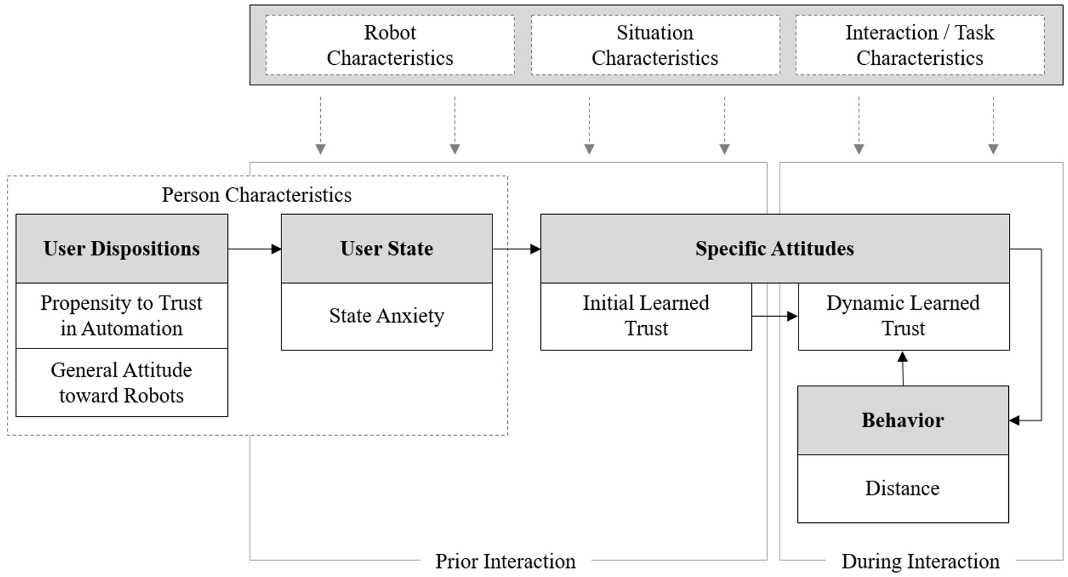


FIGURE 1 | Study framework based on the Three Stages of Trust (Kraus, 2020).

Traits, Attitudes, and Trust

Research on the influence of user characteristics on trust in HRI is relatively scarce (Hancock et al., 2011; Lewis et al., 2018). The inconclusive results on the relationship between personality traits and trust in automation call for further studies with well-founded theorizing and methodological quality to gain insight into the actual influence of personality factors in HRI (see, e.g., Kraus et al., 2020a, for a discussion of the foundation of trust in automation in personality traits). For this, a conceptual distinction between dispositional personality traits and attitudes provides an essential starting point for hypothesizing. According to Ajzen (2005), the trait and attitude concepts share substantial similarities (e.g., manifestation in observable responses) and differ regarding stability and focus. Following the definition of Ajzen (1989), *attitudes* can be understood as "an individual's disposition to respond favorably or unfavorably to an object, person, institution, or event" (p. 241). Whereas attitudes entail an evaluation of an object which is more prone to be changed, for example, by new information, personality *traits* refer to "response tendencies in a given domain" (Ajzen, 2005, p. 6) and are not oriented toward a specific object. Both traits and attitudes can differ in their specificity, as reflected in domain-specific traits (e.g., the propensity to trust in automation) and attitudes (e.g., attitude toward robots). A combined consideration of both technology-related traits and attitudes is valuable for explaining individual differences in HRI. Based on this reasoning, in this research, the domain-specific personality trait propensity to trust in automation is integrated along with the global attitude toward robots on the dispositional level, predicting user state, learned trust, and proxemic preferences.

Propensity to Trust in Automation

The dispositional layer of trust in automation, the *propensity to trust in automation*, can be defined as "an individual's overall tendency to trust automation, independent of context or a specific system" (Hoff and Bashir, 2015, p. 413). Following Mayer et al. (1995), the propensity to trust automation is hypothesized to influence learned trust. This relationship was supported before by empirical findings on a positive association between the dispositional propensity to trust and system-specific initial and dynamic learned trust (e.g., Merritt and Ilgen, 2008; Merritt et al., 2013; Kraus et al., 2020a). In line with this, in the domain of HRI, Tussyadiah et al. (2020) recently reported that users with a higher propensity to trust in technology in general hold more trusting beliefs toward a service robot. In addition to these findings, the presented research investigates the relationship between the three trust layers propensity to trust in automation, initial, and dynamic learned trust in a service robot. Accordingly, it was hypothesized that:

> *Hypothesis 1: The propensity to trust in automation is positively related to initial and dynamic learned trust in a robot (H1.1). The effect of propensity to trust on dynamic learned trust is mediated by initial learned trust (H1.2).*

Negative Attitude Toward Robots

Besides the propensity to trust, the predictive power of domain-specific attitudes for trust formation was supported in previous research. In automated driving, the prior attitude toward automated driving systems was linked to trust in automation in several studies (Singh et al., 1993; Merritt et al., 2013; Kraus et al., 2020a). Also, in the robotic domain, the role of (negative) *attitudes toward robots* was investigated before (e.g., Nomura et al., 2006a,b; Syrdal et al., 2009; Tsui et al., 2010). Thereby, previous research underlined the role of robot attitudes for the evaluation and interaction with robots (e.g., Nomura et al., 2007; Cramer et al., 2009; Syrdal et al., 2009; De Graaf and Allouch, 2013b). Yet, studies on the relationship between attitudes toward robots and trust layers arrived at inconclusive results. While Sanders et al. (2016) could not find an association between implicit attitudes and dynamic learned trust, Wang et al. (2010) reported lower dynamic learned trust to be associated with a more negative attitude toward robots. In line with the latter, Tussyadiah et al. (2020) reported a strong negative association between a general negative attitude and trusting beliefs in a service robot. While these mixed findings might be in part resulting from conceptual underspecifications of both the included attitude and trust variables, in this research in line with the proposed conceptual differentiation of trust layers and between traits and attitudes, a relationship between prior domain attitudes and learned trust was hypothesized:

Hypothesis 2: Negative attitudes toward robots are negatively related to initial and dynamic learned trust in a robot (H2.1). The effect of negative attitudes toward robots on dynamic learned trust is mediated by initial learned trust (H2.2).

Furthermore, the degree of experience and familiarity with an interaction partner is expected to influence users' trust levels. The more often one interacts with a partner, the better and more realistically the trustworthiness can be evaluated and aligned with one's own experiences. This is supported by numerous findings regarding trust in automation in general and trust in HRI in particular, which show trust to increase over time with repeated error-free interaction and growing familiarity (e.g., Muir and Moray, 1996; Beggiato and Krems, 2013; van Maris et al., 2017; Yu et al., 2017; Kraus et al., 2019b). Therefore, it is hypothesized that:

Hypothesis 3: Learned trust in a robot increases with repeated error-free interaction.

The Role of State Anxiety for Trust in Robots

Besides user dispositions, users' emotional states during the familiarization with a robot are a potential source of variance for robot trust. As the experience of emotional states has been shown to be considerably affected by personal dispositions, this research proposes a general mediation mechanism from the effects of user dispositions on trust in automation by user states (see **Figure 1**).

The presented research focuses on *state anxiety* as a specific affective state, which is expected to explain interindividual differences in trust in robots (e.g., Nomura et al., 2007; Kraus et al., 2020b). State anxiety is defined as "subjective, consciously perceived feelings of apprehension and tension, accompanied by or associated with activation or arousal of the autonomic nervous system" (Spielberger, 1966, p. 17). It is posited to "initiate a behavior sequence designed to avoid the anger situation or [...] evoke defensive maneuvers which alter the cognitive appraisal of the situation" (Spielberger, 1966, p. 17). Thereby, state anxiety was found to selectively direct attention to anxiety-triggering stimuli (Mathews and MacLeod, 1985; MacLeod and Mathews, 1988). Following the reasoning of the affect-as-information approach, users who encounter a robot for the first time might use their emotional states to build their trust toward the unfamiliar technology.

Regarding trust in interpersonal relationships, emotional states were found to influence a person's trust level (Jones and George, 1998; Dunn and Schweitzer, 2005; Forgas and East, 2008). For example, the results from Dunn and Schweitzer (2005) indicate that positive emotional states (e.g., happiness) positively and negative emotional states (e.g., anger) negatively affect trust in an unfamiliar trustee. Moreover, affective states were found to be related to trust in different automated systems (e.g., state anxiety, Kraus et al., 2020b; positive and negative affect, Stokes et al., 2010; Merritt, 2011). Interestingly, Stokes et al. (2010) found that affect was especially relevant in early trust formation processes. The relative influence diminished after repeated interaction and was replaced by more performance-related cues. These findings are in line with Lee and See's (2004) assumptions, who claimed initial trust levels to follow affective processing, and subsequent trust to be guided more by analytical processes (including, e.g., perceptions of system performance). Accordingly, the results of Kraus et al. (2020b) indicate that state anxiety predicts trust differences. Thereby, anxiety was a stronger predictor for trust in automation than negative and positive affect. As throughout the interaction more specific and tangible information about the robot becomes available, it is further hypothesized that the effects of emotional states on the actual level of learned trust diminish. Taken together, it was hypothesized:

Hypothesis 4: State anxiety is negatively related to initial learned trust in a robot (H4.1). The relationship between state anxiety and learned trust in a robot diminishes with repeated interaction (H4.2).

In the interpersonal context, people with low dispositional trust are assumed to expect others to be dishonest and potentially dangerous (Gurtman, 1992; Mooradian et al., 2006). As anxiety might facilitate oversensitivity and overinterpretation of potential threats and risk, it might mediate the link of the propensity to trust and learned trust. This assumption is supported by the findings of Kraus et al. (2020b), which show a mediation effect of state anxiety between several personality traits and dynamic learned trust in the interaction with an automated driving system. Taken together, it was expected that:

Hypothesis 5.1: The relationship between the propensity to trust in automation and initial learned trust in a robot is mediated by state anxiety (H5.1).

Above this, it is assumed that attitudes toward robots reflect the overall evaluation emerging, for example, from the assessment of their utilitarian and hedonistic benefits (Brown and Venkatesh, 2005; De Graaf and Allouch, 2013a). Several studies reported a high positive correlation between negative attitudes toward robots and state anxiety (e.g., Nomura et al., 2006b, 2008). Based on this, a mediation of the effect of negative robot attitudes on trust in a specific robot by state anxiety was hypothesized as follows:

Hypothesis 5.2: The relationship between negative attitudes toward robots and initial learned trust in a robot is mediated by state anxiety (H5.2).

Trust and Distancing Behavior

In this study, interindividual comfort zones toward robots were investigated, whereas distancing behavior was adopted as an objective interaction behavior. Spatial proximity is an essential part of human relationships. People prefer to maintain a personal space around themselves, which is expected not to be violated by others (see Hayduk, 1978, for an overview). A violation of the personal space may lead to the experience of threat and discomfort (Hayduk, 1978; Perry et al., 2013). Therefore, robot proxemic behavior design is vital for establishing close relationships and comfortable collaborations between humans and robots.

To explain distancing behavior and its function, different approaches of human relations can be drawn on (see Leichtmann and Nitsch, 2020, for an overview). Following the affect-as-information approach, an *arousal-regulating function* can be ascribed to interpersonal distancing behavior to prevent an information overload and maintain a balanced arousal level (Aiello, 1987; Leichtmann and Nitsch, 2020). Hence, arousal models argue for the change of arousal level due to interaction and approach, which leads to a cognitive evaluation and behavioral adaption (Aiello, 1987). In line with this, personal spaces can be seen as a function of perceived threat, with external sources of threat causing larger distances (anxiety-defense process, Meisels and Dosey, 1971). Regarding the interaction with robots, distancing behavior's arousal-regulatory function could play a significant role in the first encounter with this unfamiliar sophisticated technology to reduce unpleasant affective states. In this research, it is assumed that people will adapt their comfort zone toward the robot based on their initial trust level.

As theorized in different models on trust in automation (Lee and See, 2004; Hoff and Bashir, 2015; Kraus, 2020), trust has been found to be a major antecedent of reliance and, thus, of behavioral outcomes in the interaction with various technological systems (Lee and Moray, 1994; Muir and Moray, 1996; Lewandowsky et al., 2000; Hergeth et al., 2016; Payre et al., 2016). In line with this, the study by Babel et al. (2021) found a strong negative correlation between spatial distance and initial trust in a robot in a human-robot approach paradigm. The underlying mechanism of this association might be a feedback process (as proposed in the investigated research model, **Figure 1**), in which the repeated interaction with a robot might lead to an adjustment of trust based on previous interaction outcomes. In this regard, the experience made in a certain distance from the robot might influence trust, which in turn informs subsequent proximity decisions. Such a feedback process was supported by findings from MacArthur et al. (2017). In the same manner, Kraus et al. (2019b) reported a dynamic adaption of trust over the course of interaction to changing circumstances such as system malfunction. Deduced from this, it was hypothesized that:

> *Hypothesis 6: Initial learned trust and dynamic learned trust in a robot are negatively related to the comfort distance toward a robot.*

Similar to trust, experience and familiarity with the interaction partner are assumed to influence proximity preferences positively. It is to be expected that people will interact in closer proximity with trusted than with untrusted partners (e.g., closer interaction with family and friends compared to strangers). In line with this, habituation effects were found by several authors concerning allowable distances, showing decreasing distances between people and robots with growing experience and familiarity (e.g., Koay et al., 2007; Haring et al., 2013; Lauckner et al., 2014). Therefore, it is hypothesized that:

> *Hypothesis 7: Comfort distance toward a robot decreases with repeated error-free interaction.*

The proposed hypotheses were investigated in a laboratory study, in which lay users encountered and interacted with a domestic service robot for the first time in real life. In the following, the study methods, procedure, and design are depicted in more detail.

MATERIALS AND METHODS

In the presented laboratory experiment a trait-state-behavior mediation cascade in initial encounters with a domestic service robot was investigated to enhance the understanding of psychological mechanisms of trust formation and associated proximity preferences. Namely, the influences of user characteristics (dispositions and affective state anxiety) on learned trust and distancing behavior were analyzed in a repeated measure design. After the first familiarization with a humanoid robot, participants were approached by the robot two times and indicated their trust and comfort distance for each trial.

Sample Characteristics

Participants were invited to meet a domestic robot for home assistance for the first time. They had to be fluent in German and be 18 years or above. In total, 34 participants took part in the study. After the exclusion of six participants (technical issues with the robot, three times; non-compliance with instruction, one time; univariate statistical outlier regarding distance, two times), the final sample for this study consisted of $N = 28$ participants (16 female) with an average age of $M = 30.32$ ($SD = 13.61$, ranging from 18 to 60 years). Participants' technical affinity (scale of Karrer et al., 2009) was in an above-average range with $M = 4.90$ ($SD = 1.21$), trait anxiety (scale of Spielberger et al., 1970; Laux et al., 1981) on a rather medium level with $M = 2.98$ ($SD = 0.81$), both on a scale from 1 (totally disagree) to 7 (totally agree). Half of the sample indicated to be experienced with robots, which included industrial robots ($n = 5$) or vacuum cleaning robots ($n = 10$). Seven participants indicated owning a vacuum cleaning robot. None of the participants reported any personal experience or interactions with a humanoid service robot.

Experimental Setup

While this study focuses on the correlative findings of the study, in the original design also an experimental manipulation was included, which is not part of this research but will now be described to ensure a complete picture of the study setup. In the laboratory experiment in a 2 × 2-mixed design, the manipulator outreach of a humanoid robot (TIAGo, see **Figure 2**) as a between-subject factor (retracted vs. extended; $n_{retracted}$ = 14, $n_{extracted}$ = 14) and the size of the robot as a within-subject factor (short vs. tall) were manipulated. In the extracted manipulator condition, the robot stretched out his arm at a right angle toward the participant, leading to a minimum distance between the participant and the robot of 0.63 m (measured from end of the mobile base to end-effector). In two trials, the robot drove toward the participant either in the short (1.10 m) or tall (1.45 m) height condition, which was randomized in order. Therefore, within the respective groups, half of the subjects faced the short robot first, the other half the tall robot. Considering the presentation order, four experimental counterbalanced groups resulted (n = 7 each). However, the results of the experimental manipulation are not part of this report and are described elsewhere (Miller et al., 2021).

Procedure

Figure 3 provides an overview of the overall study procedure. Participants were invited to meet a robotic housekeeping assistant for personal use. To account for the different trust layers and the timely sequence of the assumed relationships, initial learned trust was measured one time in the experimental scenario (t_0), while dynamic learned trust and the distance preference were measured repeatedly (t_1, t_2). **Figure 3** details how the different trust layers were successively addressed and measured in the study.

After providing informed consent, unboxing, and watching a short demonstration of the robot's capabilities, participants took part in a semi-structured interview. After this, participants answered personality questionnaires and indicated their initial learned trust and state anxiety (t_0). To check for the influence of subjective robot perceptions, different robot evaluations were assessed in advance as control variables. **Table 1** provides an overview of the bilateral correlations of the robot evaluations, examined user charactersistics and dependent measurements. The experimental groups did not differ significantly in any of the listed measures.

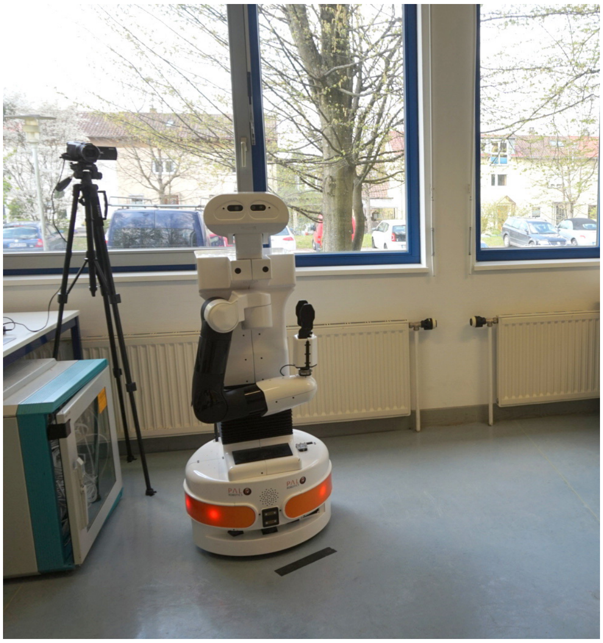


FIGURE 2 | Investigated humanoid robot (TIAGo, PAL Robotics) in initial position with retracted manipulator.

For the part of the study, which is of interest here, the robot drove toward the participants (robot-human approach) two times. This was implemented with a Wizard of Oz paradigm, in which an operator remotely controlled and stopped the robot. In accordance with the stop-distance-technique (Hayduk, 1978) similarly applied in previous proximity studies in HRI (e.g., Koay et al., 2007; Syrdal et al., 2007), participants were instructed to stand on a marked spot in 3.70 m distance to the robot. They were asked to say "stop" as soon as they started to feel uncomfortable and wanted the robot to stop approaching (comfort distance; Leichtmann and Nitsch, 2020). After each trial, the experimenter measured the spatial distance and the participants indicated their dynamic learned trust (t_1 and t_2). The second trial immediately followed the first one. At the end of the study, participants answered questionnaires, including demographic variables and their previous experience with robots. In total, the study lasted around 45–60 min.

Materials

Robot

For the study the humanoid robot TIAGo (PAL Robotics, see **Figure 2**) was used, which is variable in height (0.35 m torso lift, between 1.10 and 1.45 m) and has 12 degrees of freedom, e.g., it can move his head, arm (0.87 m reach), torso and mobile base. The robot was chosen to create consistency between the robot's appearance and its application in a private domestic environment (e.g., Lohse et al., 2008). The maximum speed of the robot is 1 m/s, by which it approached the participants.

User Dispositions

The *propensity to trust in automation* was measured with four items of the Propensity to Trust Scale (Merritt et al., 2013). The adopted translation of Kraus et al. (2020a) shows high reliability (α = 0.91) and refers to "automated technology" instead of "machines" (e.g., "I usually trust automated technology until there is a reason not to."). *Attitudes toward robots* were assessed with a self-translated eight-item version of the Negative Attitude toward Robots Scale (NARS) by Nomura et al. (2006c; e.g., "I would feel uneasy if robots really had emotions."), which assesses humans' overall attitude toward communicating robots. An English translation has shown acceptable psychometric quality (Tsui et al., 2010). Within this research, no differentiation between subscales was made, but an overall rating of the whole scale was used due to the ambiguous findings on the cultural fairness of the original loading structure (Syrdal et al., 2009; Tsui et al., 2010).

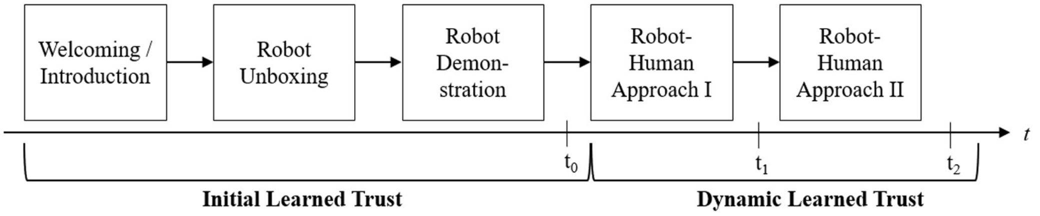


FIGURE 3 | Study procedure of the laboratory study and times of measurement including the two layers of learned trust.

TABLE 1 | Descriptives of a priori robot evaluations and bivariate correlations with user dispositions, state anxiety and trust layers at the different times of measurement.

	M	*SD*	PTT	NARS	SA	ILT	DLT t_1	DLT t_2	DT t_1	DT t_2
Competence	4.55	1.16	0.36	0.17	**−0.47**	**0.41**	0.19	0.18	0.16	0.32
Anthropomorphism	3.24	1.14	0.24	−0.09	−0.07	0.37	0.16	0.15	0.23	0.15
Uncanniness	3.18	1.19	−0.33	**0.46**	**0.41**	**−0.63**	**−0.62**	**−0.62**	−0.16	−0.20

Significant correlations in bold ($p < 0.05$, two-sided test).
PTT, Propensity to Trust in Automation; NARS, Negative Attitude toward Robots; SA, State Anxiety; ILT, Initial Learned Trust; DLT, Dynamic Learned Trust; DT, Distance.

State Anxiety

The German short version of the State-Trait Inventory (STAI; Spielberger et al., 1970) translated by Laux et al. (1981) was used to assess the participants' *state anxiety*. The state-scale (STAI-S) measures the cognitive and emotional components of anxiety as a state with five negatively and five positively poled items (e.g., "I feel tense.," "I am calm."). The STAI is considered a standard instrument in anxiety and stress research and shows high psychometric quality standards (Spielberger et al., 1970).

Learned Trust

Initial learned trust and *dynamic learned trust* were assessed as an unidimensional variable with the seven-item LETRAS-G (Kraus, 2020). Previous studies reported a high reliability of the scale (e.g., Kraus et al., 2019b). The items of the LETRAS-G were adapted to refer to "robots" instead of "automation" (e.g., "I trust the robot.").

Distance

The robot's spatial distance was measured in meters from the end of the robot's mobile base to the subject's toe after each trial. In the extended manipulator condition, 0.63 m (manipulator reach) were subtracted from the distance measure so that the value refers to the distance between the end-effector and the subject's toe.

Except for the comfort distance, all constructs were assessed using self-report short-scales. All scales were measured with a 7-point Likert scale (1 = totally disagree, 7 = totally agree). All Cronbach's α (see **Table 2**) were in an acceptable high range of ≥0.70 (Ullman, 2013), except the scale assessing the negative attitude toward robots (α = 0.67).

Statistical Procedure

To test the study hypotheses, scale means were calculated and used for all statistical procedures. The relationships of user dispositions and state anxiety with learned trust and comfort distance were calculated using regression and mediation analysis. Bivariate relationships were tested with the Pearson product-moment correlation. The reported results refer to a one-sided test in the case of directed hypotheses. Changes through repeated interaction were assessed with paired *t*-tests or ANOVAs. All analyses were conducted in R, version 3.6.2. For mediation analysis, the R package *mediation* version 4.5.0 was used as described by Tingley et al. (2014). The mediation effect (indirect effect) was tested with the bootstrapped 95% confidence intervals (CIs) with 5,000 samples (e.g., Hayes, 2018).

Preconditions

Regarding preconditions for the applied methods, first, there was no missing data in the overall data set. Second, an outlier analysis along the mean values indicated that the distance measurement of two subjects in the second trial exceeded a z-score of |3.29| (Ullman, 2013; distance = 1.10 m and 1.04 m, Mdn = 0.19 m, IQR = 0.21 m). As mentioned above, these two subjects were excluded from further analysis. All other values did not show any outliers. Third, Shapiro-Wilk tests indicated no significant deviations from a normal distribution for all mean values in the overall sample (all $p > 0.01$). Fourth, the effects of the experimental manipulation of the robot's appearance (size and manipulator outreach) on the relevant constructs in this study were analyzed using general linear models. The results showed no effects of the experimental manipulations on dynamic learned trust for neither trial (t_1 and t_2). On the other hand, the size of the robot had a significant effect on the comfort distance in the first trial, $b = 27.57$, $t(24) = 2.33$, $p = 0.028$, with the distance being larger in the tall robot ($M_{tall} = 0.47$ m, $SD_{tall} = 0.18$ m) than the short robot condition ($M_{short} = 0.23$ m, $SD_{short} = 0.25$ m). Thus, the results regarding the hypotheses on relationships with the comfort distance for t_1 should be interpreted taking the effects of the robots' size manipulation into account. Furthermore, to rule out biases due to group effects, a series of general linear models was run for each user disposition to check for interactions with the experimental manipulations on the dependent measures. The results showed no significant interactions on dynamic learned trust for neither trial (t_1 and t_2).

TABLE 2 | Means, standard deviations, correlations, and reliabilities of the included variables.

		M	*SD*	1	2	3	4	5	6	7	8
1	Propensity to trust	4.97	1.21	(0.70)							
2	Negative attitude	3.40	0.94	−0.12	(0.67)						
3	State anxiety	2.98	0.98	**−0.42**	0.01	(0.76)					
4	Initial learned trust	5.10	0.96	**0.42**	**−0.47**	**−0.59**	(0.85)				
5	Dynamic learned trust t_1	5.34	0.75	**0.40**	**−0.58**	−0.36	**0.79**	(0.81)			
6	Dynamic learned trust t_2	5.54	0.84	0.20	**−0.57**	−0.31	**0.72**	**0.85**	(0.87)		
7	Distance t_1	34.82	24.72	0.10	0.20	0.08	−0.03	0.01	0.26	–	
8	Distance t_2	18.64	13.79	−0.19	0.30	−0.01	−0.19	−0.04	0.17	**0.58**	–

Diagonal: Scale reliabilities (Cronbach's α). Significant correlations in bold (p < 0.05, two-sided test).

RESULTS

Table 2 provides the mean values, standard deviations, reliability, and correlations for all included scales for the complete cleaned sample. Due to the relatively small sample size, only correlations of $r = 0.40$ or higher reached a significant p-value ($p < 0.05$) with a two-sided test. According to Cohen (1988), correlation coefficients above $r = 0.30$ are considered as moderate effects. While we do not interpret non-significant results, the effect size might be considered as a preliminary indication of the existence of the respective relationship in the population.

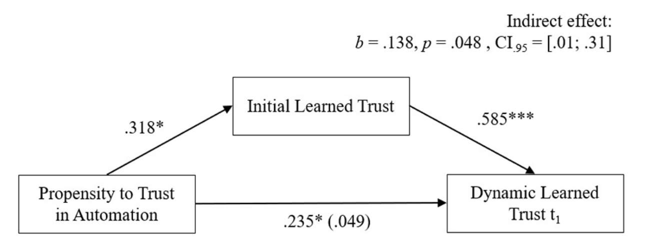

Indirect effect:
$b = .138, p = .048$, CI$_{.95}$ = [.01; .31]
Initial Learned Trust
.318*
.585***
Propensity to Trust in Automation
Dynamic Learned Trust t_1
.235* (.049)

FIGURE 4 | Mediation model for the three investigated trust layers.

Interrelation of Trust Layers

The investigated research model proposes positive relationships between the propensity to trust with initial and dynamic learned trust (H1.1). Specifically, a mediation effect from the propensity to trust over initial learned trust on dynamic learned trust was hypothesized (H1.2) to examine the trust formation process and timely sequence of the different trust layers. Accordingly, it was expected that participants would trust the robot more with growing familiarity and experience. Therefore, learned trust in the robot was expected to increase with repeated interaction (H3). Drawn from the correlation coefficients (see **Table 2**), the propensity to trust was found to be significantly related to initial learned trust ($r = 0.42$, $p = 0.014$) and dynamic learned trust in the first ($r = 0.40$, $p = 0.018$) but not in the second trial ($r = 0.20$, $p = 0.158$). As can be seen in **Figure 4**, the effect of the propensity to trust on dynamic learned trust in the first trial was fully mediated by initial learned trust, which was substantiated by the significant statistical test of the mediation effect, $b = 0.138$, $p = 0.048$, CI$_{0.95}$ = [0.01;0.31]. For the development of dynamic learned trust over time, the results of a repeated measure ANOVA supported H3, and indicated a significant linear trend, $F(1, 27) = 11.50$, $p = 0.002$, $\eta^2 = 0.299$, as reflected in increasing means of learned trust at the different points of measurement (regardless of the experimental manipulations): initial learned trust before the interaction ($M_{t0} = 5.10$, $SD_{t0} = 0.96$), dynamic learned trust in the first ($M_{t1} = 5.34$, $SD_{t1} = 0.75$) and second trial ($M_{t2} = 5.54$, $SD_{t2} = 0.84$). These findings support a dynamic increase of learned trust toward a specific robot with emerging familiarity and interaction. Considering the previous assumptions and results, H1.1 and H1.2 can be partly accepted for dynamic learned trust in the first trial and H3 can be fully accepted.

Effects of User Dispositions and State Anxiety on Initial Learned Trust

Hypothesis 2 stated that user attitudes influence learned trust in a robot. Besides the propensity to trust, an effect from negative attitudes toward robots on initial and dynamic learned trust was expected (H2.1). Similar to the mediation effect for the different trust layers, the effect of negative attitudes on dynamic learned trust was hypothesized to be mediated by initial learned trust (H2.2). In favor of H2.1, the correlation coefficients showed significant medium to high negative correlations between negative attitudes toward robots with initial learned trust ($r = -0.47$, $p = 0.006$), dynamic learned trust after the first ($r = -0.58$, $p < 0.001$) and the second trial ($r = -0.57$, $p < 0.001$). In line with H2.2, a mediation analysis showed a significant indirect effect of negative attitudes over initial learned trust on dynamic learned trust in the first trial ($b = -0.236$, $p < 0.001$, CI$_{0.95}$ = [−0.41; −0.09]; see **Figure 5**).

Above this, the results of the regression analysis with both dispositional predictors revealed a significant influence of both propensity to trust, $\beta = 0.350$, $t(25) = 2.24$, $p = 0.034$, and negative attitude toward robots, $\beta = -0.408$, $t(25) = -2.61$, $p = 0.015$, on initial learned trust. The model accounted for 29.78% of variance in the criterion.

Regarding initial learned trust, it was furthermore assumed that people who feel more anxious in anticipation of an imminent interaction with a robot would have less trust in the robot before initially interacting with it (H4.1). Furthermore, this research's overall theoretical assumption was that affect before a direct

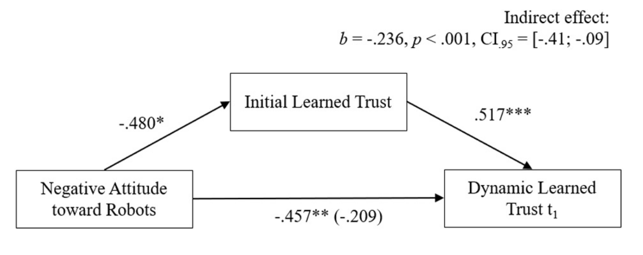

Indirect effect:
b = -.236, p < .001, CI$_{.95}$ = [-.41; -.09]
Initial Learned Trust
-.480*
.517***
Negative Attitude toward Robots
Dynamic Learned Trust t$_1$
-.457** (-.209)

FIGURE 5 | Findings for the mediation from negative attitudes toward robots to dynamic learned trust by initial learned trust.

interaction is related to the initial trust level. In turn, the latter is proposed to constitute a basis for dynamic learned trust in the early interaction with a robot. This influence of initial learned trust on dynamic learned trust is assumed to be replaced by more performance-related new information in subsequent ongoing interaction with a robot. Therefore, a decreasing correlation strength was expected from initial learned to dynamic learned trust and from the first to the second trial (H4.2). In accordance with H4.1 and H4.2, the correlation coefficients between state anxiety and learned trust showed a decrease over time, with a highly significant negative correlation between state anxiety and initial learned trust ($r = -0.59$, $p < 0.001$), and a medium significant negative correlation with dynamic learned trust in the first trial ($r = -0.36$, $p = 0.031$), and a non-significant negative correlation in the second trial ($r = -0.31$, $p = 0.054$). The results thus supported H4.

As additionally assumed in H5.1 and H5.2, the effect of the two dispositions on initial learned trust was expected to be mediated by the current affective state in the situation. To test these effects, parallel-mediation analyses were computed with the respective disposition as predictor, initial learned trust as criterion and state anxiety as mediator. In line with H5.1, the statistical test of the mediation effects (see **Figure 6**) showed a full mediation from propensity to trust on initial learned trust through state anxiety, indicated by a significant indirect effect, $b = 0.159$, $p = 0.020$, CI$_{0.95}$ = [0.02;0.39]. Due to the non-significant relationship between negative attitude toward robots as predictor and state anxiety as mediator ($r = 0.01$, $p = 0.472$), which is a prerequisite for calculating mediation analysis, no mediation for this effect was tested. Thus, in this research H5.2 could not be supported.

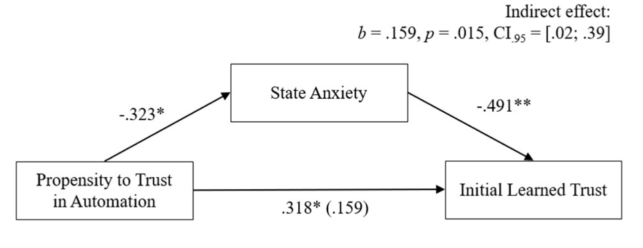

Indirect effect:
b = .159, p = .015, CI$_{.95}$ = [.02; .39]
State Anxiety
-.323*
-.491**
Propensity to Trust in Automation
Initial Learned Trust
.318* (.159)

FIGURE 6 | Findings for the mediation from the propensity to trust to initial learned trust by state anxiety.

Effects of Learned Trust on Distancing Behavior

Hypothesis 6 proposed that people with lower trust in the robot prefer to keep more distance when being approached by the robot. Dynamic learned trust and the comfort distance were measured repeatedly at t_1 and t_2. As shown in the correlation matrix (**Table 2**), the comfort distance at neither t_1 nor t_2 was significantly related to any of the remaining variables. Therefore, H6 has to be rejected.

Besides, in accordance with the findings of trust increase over time, it was assumed that participants would let the robot approach closer with growing familiarity and experience. Therefore, the comfort distance toward the robot was expected to decrease from the first to the second trial (H7). Participants in fact allowed the robot to come closer with repeated interaction (M_{t1} = 0.35 m, SD_{t1} = 0.25 m; M_{t2} = 0.19 m, SD_{t2} = 0.14 m). The results of a paired t-test showed a significant result, $t(27) = 4.24$, $p < 0.001$, *Cohen's d* = 0.808. Therefore, an overall main effect was supported by the reported findings. H7 can thus be accepted.

DISCUSSION

This research investigated the early trust development toward an unfamiliar service robot in a domestic environment prior to and during initial familiarization. Especially, the interrelation of different layers of trust was investigated, as well as the foundation of differences in learned trust in the robot in both user dispositions and affective user states. Furthermore, the role of these variables for interindividual differences in proximity preferences was investigated.

Summary of Results

Taken together, the results of the study supported eight of the eleven study (sub-)hypotheses. First of all, in line with the investigated research model and the assumption of different trust layers, the dispositional propensity to trust was positively related to the two layers of initial and dynamic learned trust (H1.1). Furthermore, the relationship between propensity to trust and dynamic learned trust was mediated by initial learned trust in the robot in the first trial (H1.2). Besides the propensity to trust, negative attitudes toward robots were negatively related to both initial and dynamic learned trust (H2.1). Similarly, the relationship between negative attitudes and dynamic learned trust was mediated by initial learned trust in the robot (H2.2). The results emphasize that domain-specific user dispositions, to some extent, influence trust ratings in the early interaction with unfamiliar technologies. The importance of this embeddedness of specific trust in user dispositions is especially emphasized by the large proportions of variance in initial learned trust, explained by the propensity to trust in automation and negative robot attitudes. Above this, in accordance with H3, learned trust in the robot increased throughout the experiment. The finding underlines that a familiarization effect takes place relatively quickly, and that lay users can get used to domestic robots after a short period already. This was further supported by a decreasing distance participants kept with repeated trials (H7).

Besides the formation of trust with emerging familiarity, the findings underline the notion that learned trust in a robot is affected by the experience of anxiety prior to the interaction with a robot. Thereby, a declining strength of the relationship between state anxiety and trust over time was found. While initial learned trust was strongly affected by the initial level of anxiety (H4.1), dynamic learned trust after the two trials showed diminished relationships with state anxiety (H4.2). Most interestingly, the relationship between the general propensity to trust in automated technology and initial learned trust in the robot was mediated by the user's initial level of state anxiety (H5.1). In contrast, no similar mediation effect for negative attitude toward robots was found (H5.2). Finally, the comfort distance toward the robot was not correlated with any of the investigated psychological constructs, contradicting H6.

Overall, the findings highlight the role of personality differences and individual variances in the technology-related learning history for the experience of anxiety and the formation of trust in automation. On this basis, a consideration of these findings in robot development and design can favor positive interaction outcomes such as safe usage, appropriate trust, and comfortable interaction. Before illustrating practical implications, this research's theoretical contributions are discussed in more detail.

Interrelation of Different Trust Layers

The reported findings underline the relevance of different layers of trust in automation, as proposed, for example, by Hoff and Bashir (2015). The findings demonstrated that inter- and intraindividual trust variations originate from both individual trait differences in the tendency to trust automation (propensity to trust in automation) and in the trust-related learning process prior to and during the interaction with a robot (initial and dynamic learned trust). In line with the propositions of the Three Stages of Trust framework (Kraus, 2020), this interplay of different trust layers in adopting a formerly unfamiliar robot is supported by the mediation cascade from the propensity to trust in automation via initial learned to dynamic learned trust. In accordance with described relationships between dispositional and system-specific trust in other domains (e.g., Lee and Turban, 2001; Merritt and Ilgen, 2008; Merritt et al., 2013; Kraus et al., 2020a), this mediation supports a timely order of these three trust layers throughout the familiarization process with an automated system like the investigated robot. The propensity to trust as a technology-specific personality trait reflects the sum of learning experiences with automated technology and considerably determines the expectations with which an individual enters the familiarization with a newly introduced system. Based on this personality variable, the available information prior to the first interaction with the system under consideration is used to build up a level of initial trust, which in turn builds the starting point for the trust calibration process during the actual interaction. Taken together, this research supports the notion of user dispositions and different trust layers that build onto each other in the emergence of a specific level of dynamic learned trust at a given time during the interaction with a robot.

Above this, in line with earlier research in HRI (e.g., van Maris et al., 2017; Yu et al., 2017) and the interaction with other automated technology, for example, plant simulations (Lee and Moray, 1994) and automated driving (e.g., Beggiato et al., 2015; Hergeth et al., 2015; Kraus et al., 2019b), in this study, trust in the robot was found to increase throughout the interaction incrementally. As long as there is no negative information like an experience of restricted reliability (e.g., automation malfunction; Kraus et al., 2019b), a violation of initial expectations, or realization of initial concerns and fears, accumulated positive information and experiences lead to an increase in trust over time. At the same time, this shows that despite the discussed differences between systems from different technological domains, general results from other domains might be transferable to HRI.

Derived from that, researchers must consider carefully which trust layer and which points in time are addressed in their experimental design. Notwithstanding, trust should be measured several times throughout HRI research. Furthermore, a combined consideration of dispositional trust and learned trust should be entailed in research designs. Taken together, the investigation of factors affecting the process of trust formation and calibration, in which these three trust layers build on each other, is an essential prerequisite for predicting, understanding, and modifying the interaction with a robot at a given point in time.

Role of User Dispositions and States for Trust in Human-Robot Interaction

This research identified two user dispositions and the emotional state anxiety to affect trust processes, answering the call for a more thorough investigation of user characteristics' influence in HRI (e.g., Hancock et al., 2011). Thereby, the findings go beyond previous research and emphasizes the meaningfulness of (technology-)specific personality traits and attitudes in the individual reaction to technology.

The presented study supports a relationship between state anxiety and learned trust in service robots in line with the affect-as-information approach. Robots are a new technology and most people (our sample in particular) did not have many opportunities to establish first-hand experiences. Therefore, it is not surprising that anxiety plays an essential role in the initial familiarization process with robots in the face of the associated uncertainty (and maybe also pre-existing reservations). Besides the actual anxiety, which is directly triggered by the new, unpredictable robot, also misattributions of affective states might influence trust and other evaluative outcomes. The study findings corroborate the results of studies on human interactions, in which mood-congruent judgments of trust in a co-worker (Dunn and Schweitzer, 2005) or general life satisfaction (Schwarz and Clore, 1983) were found. Interestingly, in Schwarz and Clore's (1983) work, participants only (mis)attributed their bad mood onto judgments about their lives, when no alternative (external) transient source for attributing the bad mood to was salient. In light of the presented findings, the robot offers a plausible source for the attribution of current feelings (and is at the same time one cause for these; see also in section "Reflections on

Trust Formation and Calibration Processes at Different Points in Time"). In line with other research (Stokes et al., 2010; Merritt, 2011), the results emphasize the need to consider affective states and mechanisms of information processing, especially in early interactions with new technology when the user has not yet had any experiences of his own with the system to build trust on. With growing familiarity, users start to build more on personal experiences than potentially misattributed and misinterpreted inner states.

Furthermore, state anxiety was found to be predicted by the propensity to trust in automation. Also a significant trait-state mediation from the propensity to trust to initial learned trust via state anxiety was found. Overall, the reported findings on anxiety point into the direction that individual differences in the general tendency to trust shape the affective reaction to new technology which in turn influences initial trust levels. To conclude, for understanding the interindividual variances in the reaction toward a specific robot, a consideration of pre-existing individual differences in the propensity to trust in automation and the consideration of the individual learning history and affective reaction seems worthwhile.

Findings on Distancing Behavior

In light of the affect-regulating function, this research assumed the user's trust level to serve as an information and evaluation source for the comfort distance toward the robot. However, no relationship between the behavioral measurement and trust in the robot could be found. At the same time, findings indicate a decrease of the comfort distance toward the robot over time.

There are different potential explanations for these findings. First, it should be considered that the study sample was relatively small since real face to face interactions and behavior were investigated. As a result, correlations in the area of moderate effect sizes did not reach significance. Furthermore, it seems plausible that the small to medium correlations between trust and distancing behavior are mediated and moderated by interposed processes and constructs, which were not addressed in this study. Besides, this study applied a robot-human approach with participants instructed to indicate their preferred distance. A different design (e.g., human-robot approach, field observation) might have produced other results because users could adjust the actual distance more dynamically and adapt it to the interaction context, task, and shifting inner states.

Second, there might be a direct effect of robot characteristics on proximity preferences, which is not mediated by trust. A multitude of research supports that robot-related factors influence the preferred distance toward a robot. A potential direct effect of robot characteristics on distance preferences was supported in this study (see Miller et al., 2021). Besides this, the findings underline an important role of the subjective perception of robot characteristics on the initial evaluation of and reaction to robots. Specifically, this is emphasized by the correlations of both state anxiety and the investigated trust layers with different robot evaluations (especially uncanniness).

Reflections on Trust Formation and Calibration Processes at Different Points in Time

Regarding the process of trust formation and calibration over time within different phases of familiarizing with robots and other new technology, some of the presented findings are worth to be discussed in more detail. A closer inspection of the magnitude of bivariate correlations between dispositions, state anxiety, and learned trust at different times of measurement (**Table 2**) reveals an interesting pattern. On the one hand, the strengths of both the relationships of the propensity to trust and state anxiety with learned trust decreased over time. On the other hand, the correlations between (negative) attitudes and learned trust increased in their magnitude over time. Besides, the negative attitude toward robots was not related to state anxiety, implying no similar mediation effect on learned trust as for the propensity to trust. These findings point toward differential information sources and information processing mechanisms, through which trust in a specific system is built and calibrated. In the following, two possible explanations for changing information use at different times in the familiarization with automated systems are discussed, which can account for the observed patterns: changing availability of information (sources) and different processing mechanisms depending on motivation and cognitive capacities.

First, it is reasonable to assume that different kinds of information are present at different phases during the familiarization with a new technological system. Before system use, mainly information from second-hand testimonials of users, marketing, and information campaigns are available. In contrast, when interacting with a system, the actual system behavior and user interface output provide diagnostic information to assess a system's trustworthiness. Accordingly, in the early phase of getting to know an automated system (before actual system use), the available information tends to be more vague, indirect, and unspecific. On the contrary, in the later phase(s), the information results from a direct first-hand experience of the user and tends to be associated with system behavior, the current task, and environmental conditions. Therefore, it seems reasonable to assume that the relevance of different categories of trust cues (reputation-, purpose-, process-, performance-, and appearance-related; see Kraus, 2020) changes over time. At this point, the differential impact and character of available trust cues prior and during the interaction with robots have to our knowledge not been extensively investigated and, therefore, provide promising directions for future research.

The second plausible mechanism of differential information use prior and during the interaction is a result of the characteristics of information processing in attitude formation and change. This is related to the idea of different routes of information processing by Lee and See (2004; affective, analogous, analytical) and the assumptions of dual-process theories of attitude formation and change (Chen et al., 1996; Chaiken and Trope, 1999). Dual-process theories assume two routes through which information can be processed and through which change of attitudes is initiated. For example, the Elaboration-Likelihood Model (ELM; Petty and Cacioppo, 1986) proposes a central route, through which an effortful analysis of the meaning of information is conducted (bottom-up). On the contrary, the peripheral route represents an attributional process, in which surface characteristics of the information or the information source lead to the change of attitudes (top-down). An essential prediction of the ELM (and other dual-process theories) is that the motivation and the ability of the person in focus determines which route of information processing is used to which extent (Petty and Cacioppo, 1986). The relative contribution of the two processes in trust formation and calibration is expected to change according to the information available and the associated affective, cognitive, and motivational processes at play at different points in time. In support of this, Kraus et al. (2019a) found that participants used information provided before the first interaction with an automated driving system differently in building up their trust levels based on their individual expression of *need for cognition*, which reflects an individual tendency to enjoy and engage in cognitive tasks (Cacioppo et al., 1984) and thus for effortful analysis of provided information (e.g., Cacioppo et al., 1983). From this, it can be derived that, irrespective of the availability of information, other characteristics of information and, therefore, different entailed trust cues might be used in different trust formation phases. Based on these changes in information processing, it can further be assumed that the extent to which the user's self-monitoring of, for example, psychological states (e.g., workload, stress, affect) is used as a source for trust changes over individuals and time in the trust formation process. For example, if users are not motivated or have restricted capacity to reflect (which might be the case in trust processes prior to an interaction), they might more strongly build their trust on their current feeling (e.g., affective state, "How do I feel about it?") in the sense of affect-as-information (e.g., Schwarz and Clore, 1988). Similarly, in such a situation, users might tend to engage more strongly in top-down than in bottom-up processing and exemplarily base their trust formation more on already existing general attitudes (e.g., attitude toward robots in general). On the contrary, in calibrating one's trust during (repeated) system use, the motivation to correctly assess provided trust cues to use the system adequately should be drastically increased due to the associated higher risks. As a result, the influences of affect and prior attitudes on trust might be diminished in favor of actual diagnostic trustworthiness (bottom-up) information available in a situation.

The results of this study suggest such interaction of different routes of information processing and a differential role of top-down and bottom-up processes in trust formation and calibration. Findings indicate that the affective and cognitive trust processes change in their relative importance for explaining the momentary level of learned trust. While at the beginning of the interaction, trust is more strongly affected by state anxiety and the propensity to trust (which reflect top-down processes), the influence of prior negative attitudes toward robots remains in the same range. While the proportions of variance explained by anxiety reflect the affective component of trust, the negative attitudes toward robots might reflect a more cognitive evaluative trust basis. If these attitudes are of rational nature, they might also gain relevance in effortful cognitive decision-making. Taken together, the interactive role of different information processing over time in trust formation is a promising direction for further research that might essentially contribute to an understanding of trust processes and to an appropriate design of information about robots, robot appearance, and behavior, as well as HRI in practice.

Practical Implications

Since trust in robots is essential to foster safe, efficient, and comfortable interactions, the reported results provide a foundation for the derivation of design recommendations in developing service robots for domestic environments in particular.

First of all, as reflected in personality traits and technology-related attitudes, user dispositions were found to affect both the experience of anxiety in the face of an initially unfamiliar robot and the individual trust level. This underlines the importance of taking the target group and individual user into account when designing interactions with robots. Engineers and designers might carefully consider which users they develop the robot for and if a possible differentiation of user groups in terms of their personality and attitudes is reasonable. In practical settings, users with initial negative attitudes might be addressed by providing more information on the robot's potential advantages. In contrast, it is more important for users with a positive attitude to ensure that they do not overtrust the robot and overestimate its abilities by providing transparent and adequate information about the robot's limitations. Furthermore, the investigated user dispositions and the level of anxiety might be assessed to individualize the process of introducing a robot and personalizing the robot's interaction concept. In this way, first interactions with robots, for example, for novices vs. tech-savvy users, could be set up differently and adapted to individual needs. While for novice users, anxiety-reducing interaction strategies and behaviors might be appropriate, as for example, the safe planner introduced by Kulić and Croft (2005), experienced users might already be accustomed and habituated to the movements and operating modes of the robot and therefore accept closer proxemics right away. Consequently, a customized robot might be less scary, more acceptable, and efficient, and might facilitate trust calibration.

Moreover, the reported findings emphasize the importance of how robots are advertised and promoted before handed over and entering personal spaces. To facilitate a certain level of trust in the robot, demonstrating its capabilities, functioning, and potential benefits and limitations and risks should occur beforehand to minimize anxious feelings before the first interaction with the new companion. The goal should be conveying a realistic picture of potential threats so that the users can rely on facts rather than on possibly misattributed feelings of arousal and anxiety. Taken together, while consideration of characteristics and appearance of the robot in design is essential, the individual reaction to and evaluation of a robot is considerably influenced by pre-existing differences regarding user personality and the individual learning history with technology and robots. A more detailed discussion of how personality differences can be considered in user education and the design of automated systems can be found in Kraus et al. (2020a) and Kraus (2020).

Additionally, in line with the Three Stages of Trust framework by Kraus (2020), this research underlines that different trust layers are involved in the emergence of learned trust in a robot at a specific point in time. This includes the propensity to trust in automation, which constitutes a specific personality trait and is, besides others, influenced by the individual learning history with technology. Therefore, in practice, it seems relevant to understand the users' level of expertise and design the introduction process, the provided information, and the user interface accordingly. As mentioned, the introduction might be shorter and more about the technical functioning for experienced users or technological experts. At the same time, the interface might provide different information or enable other functions with growing experience.

Furthermore, while the experiences during the actual interaction with a robot are undoubtedly important for the individual level of dynamic learned trust, the level of trust established prior to the interaction determines user expectations during the interaction with a robot. Similarly, the information provided before any interaction influences how users interpret the robot's behavior during HRI. Hence, trust processes and the available information before the actual interaction (initial learned trust) with a robot need to be considered to understand how trust in a robot at a certain point during the interaction (dynamic learned trust) is established. Therefore, both researchers and designers might consider the following questions in the understanding of the formation of learned trust in a specific automated system under consideration. What image is (currently) conveyed by media reports, how is the specific robot advertised, what information is available online and what are experiences reported by family or friends? All these information sources shape a robot's evaluation and the trust level before the user and robot even met or interacted, influencing the perception, evaluation, and reaction in HRI.

Furthermore, this study's findings underline the relevance of different types of information and different psychological processes for trust at different phases of the trust formation and calibration process. While more research is necessary to gain further insights into the relative importance of different kinds of information and information processing mechanisms at these different phases, at this point, it seems essential to focus on personality and user state effects in the initial phase of trust formation. Therefore, for user education and design of HRI, individual anxiety might be addressed more at the beginning of the interaction. In contrast, general attitude-based trust formation seems relevant for trust formation throughout all phases of early interaction. Therefore it is vital that positive attitudes are promoted by emphasizing a robot's assistance potential. This could be achieved by providing transparent information about the robot's capabilities and limitations or repeated HRI with positive experiences (building a personal learning history).

Additionally, the results show that users allow a robot to come comparatively close and act in immediate proximity to the user. Since domestic robots are likely to work and collaborate in close physical proximities with users as compared to industrial robots, this finding is of high practical relevance. It allows developing and employing robots for tasks that require close collaboration between robots and humans (like, e.g., in healthcare applications). Technical restriction (of, e.g., recognition systems) and resulting perceived impairment of robot performance should be considered here (e.g., Mead and Matarić, 2015). Furthermore, individual user characteristics and preferences could also be considered in robots' proxemic design, for example, by robots recognizing whether they have already interacted with a user or not (e.g., van Oosterhout and Visser, 2008). In this regard, reactions from preceding interactions could similarly be used to adapt the robot behavior dynamically.

Strengths, Limitations, and Future Directions

On the side of this research's strengths, specifically, the real-life interaction with a (domestic) service robot in the experimental setup has to be stressed. A second advantage of the study design is the combination of subjective variables and objective behavioral outcomes. A further strength is the derivation of the study hypotheses from theorizing in trust in automation, HRI, and broader psychological domains. This facilitates the scientific accumulation of knowledge about the interaction and the design of service robots. Furthermore, the results guide practitioners with various implications for the interaction design between humans and robots regarding the promotion of calibrated trust and adequate proxemic behavior.

Nevertheless, like all research, this study has some limitations, which might be addressed in future research. First, the investigated effects of familiarization are likely to not fully play out during the observed time frame over two directly sequential interactions. A longitudinal design with actual interaction and task completion might be implemented in future studies to show the effects under consideration in the long term, like for example in Koay et al. (2007). Second, changes in state anxiety were not assessed in the experiment. In a similar manner, the consideration of different user states, such as arousal or positive and negative affect, could have strengthened the findings. In this context, the application of physiological methods is a promising approach, for example, to draw conclusions on anxiety or stress (e.g., Crossman et al., 2018). Third, as the sample was self-selected, particularly anxious people might be underrepresented in the study. Future studies could take further steps to include anxious participants. Fourth, only one personality trait and one dispositional technology attitude were included in this study. Future research could address the relationships of further dispositional variables with trust and distance preferences. For example, the role of attitudes toward technology in general and robots in particular might be considered as well as (robot) trait anxiety. Fifth, between the two trials an experimental manipulation of the robots' size and manipulator position took place, which might additionally produce interindividual variance and changes in trust and distance (see results in Miller et al., 2021). Sixth, the sample size of this study was rather small for a correlative design as reflected in restricted power. While this was an effect of the natural experimental setting of the study, the reported significant findings underline the large effect sizes of the relationships between the variables entailed in the investigated research model.

Apart from that, this research provides an integrative theoretical basis for further consideration of the role of psychological variables associated with interindividual variances in trust in robots and HRI. On this basis, further studies aiming at a more detailed understanding of the psychological process in which trust in robots is formed and calibrated might focus on additional dispositional, situational, and robot-related variables, as well as the psychological processes, in which these interact over time. This involves interactions between dispositional user characteristics and situational variables (e.g., the role of robot expertise in interpreting a robot's distancing behavior) and interactions between dispositional variables and robot design features (like the role of cooperative personality traits for preferences of different levels of robot anthropomorphism). Additionally, future research might strive to validate the established relationships in a large-scale sample. Also, the transferability of the findings to other than domestic environments could be addressed in further studies. Finally, the results of this study raise the question if users get used to robots very quickly and therefore the implications for overconfidence might be addressed in future research. Future studies may focus on an appropriate and calibrated level of trust in robots and how this can be achieved. It remains to be seen in the next years how and on what basis social norms with robots (that have been established in the interaction between humans over decades) will develop when these are increasingly integrated into personal spaces and society.

CONCLUSION

Service robots are increasingly entering public and private spaces, which will promote close and personal interactions. This research strived to further investigate the role of user characteristics in the emergence of trust and distancing behavior in HRI. Especially the a priori attitude toward robots in general and the propensity to trust in automation seem to contribute to the understanding of interindividual differences in trust in robots and therefore affect appropriate robot use. By integrating psychological antecedents of close human-robot collaborations such as personality traits, affect and trust, this research provides a foundation for designing robots and directions for future developments. A role of a mediation mechanism from user dispositions to learned trust by state anxiety was supported. Thus, this research contributes to a deeper understanding of underlying determinants for affective and behavioral reactions in close personal interaction with robots. Taken together, the reported findings support central propositions of the Three Stages of Trust framework (Kraus, 2020) in terms of the history-based psychological process in which trust in automated systems is built up and calibrated. In this regard, the reported findings argue for considering user dispositions and processes before the actual interaction with a specific robot to understand better how evaluations and decision-making regarding one particular robot are established on a psychological level. The presented research constitutes a starting point for further research on the psychological basis of trust in robots by integrating broader personality traits, robot-related traits, and trust constructs with different specification levels and foci simultaneously. This research provides a foundation for utilizing the benefits and potentials of robots more fully and successfully integrating robots into our society and everyday life.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding authors.

ETHICS STATEMENT

This study was conducted with the consent of the ethical committee of Ulm University (approval no. 95/19). The ethical approval was granted under the condition that the data protection regulations were adhered to. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LM collected the data, performed the data analysis, and led the manuscript write-up. JK generated the research questions and study method, led the study implementation, and had a substantial part in writing and editing the manuscript. FB was substantial to study conception, design, and implementation and assisted with the manuscript write-up. MB assisted with the manuscript write-up. All authors were involved in the research process.

FUNDING

This research has been conducted within the interdisciplinary research project "RobotKoop," which was funded by the German Ministry of Education and Research (Grant No. 16SV7967).

REFERENCES

Aiello, J. R. (1987). "Human spatial behavior," in *Handbook of Environmental Psychology*, eds D. Stokols and I. Altman (New York, NY: Wiley), 389–504.

Ajzen, I. (1989). "Attitude structure and behavior," in *Attitude Structure and Function*, eds A. R. Pratkanis, S. J. Breckler, and A. G. Greenwald (New Jersey, NJ: Lawrence Erlbaum), 241–274.

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211.

Ajzen, I. (2005). *Attitudes, Personality, and Behavior*, 2nd Edn. New York, NY: McGraw-Hill Education.

Babel, F., Kraus, J., Miller, L., Kraus, M., Wagner, N., Minker, W., et al. (2021). Small talk with a robot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. *Int. J. Soc. Robot.* 1–14. doi: 10.1007/s12369-020-00730-0

Beggiato, M., and Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transp. Res. Part F Traffic Psychol. Behav.* 18, 47–57. doi: 10.1016/j.trf.2012.12.006

Beggiato, M., Pereira, M., Petzoldt, T., and Krems, J. (2015). Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transp. Res. Part F Traffic Psychol. Behav.* 35, 75–84. doi: 10.1016/j.trf.2015.10.005

Brave, S., and Nass, C. (2007). "Emotion in human-computer interaction," in *Human-Computer Interaction Fundamentals*, eds A. Sears and J. A. Jacko (Boca Raton, FL: CRC Press), 53–68. doi: 10.1201/b10368-6

Brief, A. P., and Weiss, H. M. (2002). Organizational behavior: affect in the workplace. *Annu. Rev. Psychol.* 53, 279–307. doi: 10.1146/annurev.psych.53.100901.135156

Brown, S. A., and Venkatesh, V. (2005). Model of adoption of technology in households: a baseline model test and extension incorporating household life cycle. *MIS Q.* 29, 399–426. doi: 10.2307/25148690

Buss, A. H. (1989). Personality as traits. *Am. Psychol.* 44, 1378–1388. doi: 10.1037/0003-066X.44.11.1378

Cacioppo, J. T., Petty, R. E., and Kao, C. F. (1984). The efficient assessment of need for cognition. *J. Pers. Assess.* 48, 306–307. doi: 10.1207/s15327752jpa4803_13

Cacioppo, J. T., Petty, R. E., and Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *J. Pers. Soc. Psychol.* 45, 805–818. doi: 10.1037/0022-3514.45.4.805

Chaiken, S., and Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. New York, NY: Guilford Press.

Chen, S., Shechter, D., and Chaiken, S. (1996). Getting at the truth or getting along: accuracy-versus impression-motivated heuristic and systematic processing. *J. Pers. Soc. Psychol.* 71, 262–275. doi: 10.1037/0022-3514.71.2.262

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. New Jersey, NJ: Lawrence Erlbaum.

Cramer, H., Kemper, N., Amin, A., Wielinga, B., and Evers, V. (2009). 'Give me a hug': the effects of touch and autonomy on people's responses to embodied social agents. *Comput. Animat. Virtual Worlds* 20, 437–445. doi: 10.1002/cav.317

Crossman, M. K., Kazdin, A. E., and Kitt, E. R. (2018). The influence of a socially assistive robot on mood, anxiety, and arousal in children. *Prof. Psychol.* 49, 48–56. doi: 10.1037/pro0000177

De Graaf, M. M. A., and Allouch, B. S. (2013a). Exploring influencing variables for the acceptance of social robots. *Rob. Auton. Syst.* 61, 1476–1486. doi: 10.1016/j.robot.2013.07.007

De Graaf, M. M. A., and Allouch, B. S. (2013b). "The relation between people's attitude and anxiety towards robots in human-robot interaction," in *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*, (New York, NY: IEEE), 632–637. doi: 10.1109/ROMAN.2013.6628419

Dunn, J. R., and Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. *J. Pers. Soc. Psychol.* 88, 736–748. doi: 10.1037/0022-3514.88.5.736

Fishbein, M., and Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Boston, MA: Addison-Wesley.

Forgas, J. P., and East, R. (2008). On being happy and gullible: mood effects on skepticism and the detection of deception. *J. Exp. Soc. Psychol.* 44, 1362–1367. doi: 10.1016/j.jesp.2008.04.010

Forgas, J. P., and George, J. M. (2001). Affective influences on judgments and behavior in organizations: an information processing perspective. *Organ. Behav. Hum. Decis. Process.* 86, 3–34. doi: 10.1006/obhd.2001.297

Gurtman, M. B. (1992). Trust, distrust, and interpersonal problems: a circumplex analysis. *J. Pers. Soc. Psychol.* 62, 989–1002. doi: 10.1037/0022-3514.62.6.989

Hamaker, E. L., Nesselroade, J. R., and Molenaar, P. C. M. (2007). The integrated trait–state model. *J. Res. Pers.* 41, 295–315. doi: 10.1016/j.jrp.2006.04.003

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors* 53, 517–527. doi: 10.1177/0018720811417254

Haring, K. S., Matsumoto, Y., and Watanabe, K. (2013). How do people perceive and trust a lifelike robot. *Proc. World Congr. Eng. Comput. Sci.* 1, 425–430.

Hayduk, L. A. (1978). Personal space: an evaluative and orienting overview. *Psychol. Bull.* 85, 117–134. doi: 10.1037/0033-2909.85.1.117

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, 2nd Edn. New York, NY: The Guilford Press.

Hergeth, S., Lorenz, L., Krems, J. F., and Toenert, L. (2015). "Effects of take-over requests and cultural background on automation trust in highly automated driving," in *Proceedings of the 8th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Salt Lake City, UT, 331–337. doi: 10.17077/drivingassessment.1591

Hergeth, S., Lorenz, L., Vilimek, R., and Krems, J. F. (2016). Keep your scanners peeled: gaze behavior as a measure of automation trust during highly automated driving. *Hum. Factors* 58, 509–519. doi: 10.1177/0018720815625744

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 407–434. doi: 10.1177/0018720814547570

Jones, G. R., and George, J. M. (1998). The experience and evolution of trust: implications for cooperation and teamwork. *Acad. Manage Rev.* 23, 531–546. doi: 10.5465/amr.1998.926625

Karrer, K., Glaser, C., Clemens, C., and Bruder, C. (2009). "Technikaffinität erfassen – der fragebogen TA-EG," in *Der Mensch im Mittelpunkt technischer Systeme. 8. Berliner Werkstatt Mensch-Maschine-Systeme*, Vol. 8, Hrsg. L. C. Stößel and C. Clemens (Düsseldorf: VDI Verlag GmbH), 196–201.

Koay, K. L., Syrdal, D. S., Walters, M. L., and Dautenhahn, K. (2007). "Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study," in *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*, (New York, NY: IEEE), 564–569. doi: 10.1109/ROMAN.2007.4415149

Kraus, J. (2020). *Psychological Processes in the Formation and Calibration of Trust in Automation*. Doctoral dissertation, Ulm University, Ulm.

Kraus, J., Forster, Y., Hergeth, S., and Baumann, M. (2019a). Two routes to trust calibration: effects of reliability and brand information on trust in automation. *Int. J. Mob. Hum. Comput. Interact.* 11, 1–17. doi: 10.4018/IJMHCI.2019070101

Kraus, J., Scholz, D., Stiegemeier, D., and Baumann, M. (2019b). The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Hum. Factors.* 62, 718–736. doi: 10.1177/0018720819853686

Kraus, J., Scholz, D., and Baumann, M. (2020a). What's driving me? - Exploration and validation of a hierarchical personality model for trust in automated driving. *Hum. Factors* doi: 10.1177/0018720820922653 [Epub ahead of print].

Kraus, J., Scholz, D., Messner, E. M., Messner, M., and Baumann, M. (2020b). Scared to trust? – predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Front. Psychol.* 10:2917. doi: 10.3389/fpsyg.2019.02917

Kulić, D., and Croft, E. (2005). "Anxiety detection during human-robot interaction," in *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (New York, NY: IEEE), 616–621. doi: 10.1109/IROS.2005.1545012

Lauckner, M., Kobiela, F., and Manzey, D. (2014). "'Hey robot, please step back!' - exploration of a spatial threshold of comfort for human-mechanoid spatial interaction in a hallway scenario," in *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, (New York, NY: IEEE), 780–787. doi: 10.1109/ROMAN.2014.6926348

Laux, L., Glanzmann, P., Schaffner, P., and Spielberger, C. D. (1981). *Das State-Trait-Angstinventar (STAI)*. Bad Langensalza: Beltz.

Lee, J. D., and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 1243–1270. doi: 10.1080/00140139208967392

Lee, J. D., and Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *Int. J. Hum. Comput. Stud.* 40, 153–184. doi: 10.1006/ijhc.1994.1007

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50_30392

Lee, M. K., and Turban, E. (2001). A trust model for consumer internet shopping. *Int. J. Electron. Commer.* 6, 75–91. doi: 10.1080/10864415.2001.11044227

Leichtmann, B., and Nitsch, V. (2020). How much distance do humans keep toward robots? Literature review, meta-analysis, and theoretical considerations on personal space in human-robot interaction. *J. Environ. Psychol.* 68:101386. doi: 10.1016/j.jenvp.2019.101386

Lewandowsky, S., Mundy, M., and Tan, G. (2000). The dynamics of trust: comparing humans to automation. *J. Exp. Psychol.* 6, 104–123. doi: 10.1037/1076-898X.6.2.104

Lewis, M., Sycara, K., and Walker, P. (2018). "The role of trust in human-robot interaction," in *Foundations of Trusted Autonomy*, eds H. A. Abbass, J. Scholz, and D. J. Reid (Berlin: Springer), 135–159.

Lohse, M., Hegel, F., and Wrede, B. (2008). Domestic applications for social robots - an online survey on the influence of appearance and capabilities. *J. Phys. Agents* 2, 21–32. doi: 10.14198/JoPha.2008.2.2.04

MacArthur, K. R., Stowers, K., and Hancock, P. A. (2017). "). Human-robot interaction: proximity and speed- slowly back away from the robot!," in *Advances in Human Factors in Robots and Unmanned Systems*, eds P. Savage Knepshield and J. Chen (Berlin: Springer), 365–374. doi: 10.1007/978-3-319-41959-6_30

MacLeod, C., and Mathews, A. (1988). Anxiety and the allocation of attention to threat. *Q. J. Exp. Psychol. Sec. A* 40, 653–670. doi: 10.1080/14640748808402292

Maio, G. R., Haddock, G., and Verplanken, B. (2018). *The Psychology of Attitudes and Attitude Change*. Thousand Oaks, CL: Sage Publications Limited.

Mathews, A., and MacLeod, C. (1985). Selective processing of threat cues in anxiety states. *Behav. Res. Ther.* 23, 563–569. doi: 10.1016/0005-7967(85)90104-4

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 709–734. doi: 10.5465/amr.1995.9508080335

McKnight, D., Choudhury, V., and Kacmar, C. (2002). Developing and validating trust measures for e-Commerce: an integrative typology. *Inf. Syst. Res.* 13, 334–359. doi: 10.1287/isre.13.3.334.81

Mead, R., and Matarić, M. J. (2015). "Proxemics and performance: Subjective human evaluations of autonomous sociable robot distance and social signal understanding," in *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (New York, NY: IEEE), 5984–5991. doi: 10.1109/IROS.2015.7354229

Meisels, M., and Dosey, M. A. (1971). Personal space, anger-arousal, and psychological defense. *J. Pers.* 39, 333–344. doi: 10.1111/j.1467-6494.1971.tb00046.x

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Hum. Factors* 53, 356–370. doi: 10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., Lachapell, J., and Lee, D. (2013). I trust it, but i don't know why: effects of implicit attitudes toward automation on trust in an automated system. *Hum. Factors* 55, 520–534. doi: 10.1177/0018720812465081

Merritt, S. M., and Ilgen, D. R. (2008). Not all trust is created equal: dispositional and history-based trust in human-automation interactions. *Hum. Factors* 50, 194–210. doi: 10.1518/001872008X288574

Miller, L., Kraus, J., Babel, F., Messner, M., and Baumann, M. (2021). "Come closer: experimental investigation of robot' appearance on proximity, affect and trust in a domestic environment," in *Proceedings of the 64th Human Factors and Ergonomics Society Annual Meeting*, Vol. 64, (Thousand Oaks, CL: SAGE Publishing), 395–399. doi: 10.1177/1071181320641089

Mooradian, T., Renzl, B., and Matzler, K. (2006). Who trusts? Personality, trust and knowledge sharing. *Manage. Learn.* 37, 523–540. doi: 10.1177/1350507606073424

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man Mach. Stud.* 27, 527–539. doi: 10.1016/S0020-7373(87)80013-5

Muir, B. M., and Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 429–460. doi: 10.1080/00140139608964474

Nass, C., Isbister, K., and Lee, E.-J. (2000). "Truth is beauty: researching embodied conversational agents," in *Embodied Conversational Agents*, eds J. Cassell, J. Sullivan, J. Prevost, and E. Churchill (Cambridge, MA: MIT Press), 374–402.

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Nomura, T., Kanda, T., and Suzuki, T. (2006a). Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. *AI Soc.* 20, 138–150. doi: 10.1007/s00146-005-0012-7

Nomura, T., Kato, K., Kanda, T., and Suzuki, T. (2006b). "Exploratory investigation into influence of negative attitudes toward robots on human-robot interaction," in *Mobile Robots Towards New Applications*, ed. A. Lazinica (London: IntechOpen), 784–802. doi: 10.5772/4692

Nomura, T., Suzuki, T., Kanda, T., and Kato, K. (2006c). Measurement of negative attitudes toward robots. *Interact. Stud.* 7, 437–454. doi: 10.1075/is.7.3.14nom

Nomura, T., Kanda, T., Suzuki, T., and Kato, K. (2008). Prediction of human behavior in human-robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Trans. Robot.* 24, 442–451. doi: 10.1109/TRO.2007.914004

Nomura, T., Shintani, T., Fujii, K., and Hokabe, K. (2007). "Experimental investigation of relationships between anxiety, negative attitudes, and allowable distance of robots," in *Proceedings of the 2nd IASTED International Conference on Human-Computer Interaction*, (Chamonix: ACTA Press), 13–18.

Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* 39, 230–253. doi: 10.1518/001872097778543886

Payre, W., Cestac, J., and Delhomme, P. (2016). Fully automated driving: impact of trust and practice on manual control recovery. *Hum. Factors* 58, 229–241. doi: 10.1177/0018720815612319

Perry, A., Rubinsten, O., Peled, L., and Shamay-Tsoory, S. G. (2013). Don't stand so close to me: a behavioral and ERP study of preferred interpersonal distance. *NeuroImage* 83, 761–769. doi: 10.1016/j.neuroimage.2013.07.042

Petty, R. E., and Cacioppo, J. T. (1986). "The elaboration likelihood model of persuasion," in *Advances in Experimental Social Psychology*, ed. L. Berkowitz (Cambridge, CA: Academic Press), 123–205.

Robinson, H., MacDonald, B., and Broadbent, E. (2014). The role of healthcare robots for older people at home: a review. *Int. J. Soc. Robot.* 6, 575–591. doi: 10.1007/s12369-014-0242-2

Rosenthal-von der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., et al. (2014). Investigations on empathy towards humans and robots using fMRI. *Comput. Hum. Behav.* 33, 201–212. doi: 10.1016/j.chb.2014.01.004

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, (New York, NY: IEEE), 1–8. doi: 10.1145/2696454.2696497

Sanders, T. L., Schafer, K. E., Volante, W., Reardon, A., and Hancock, P. A. (2016). "Implicit attitudes toward robots," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60, (Washinton, DC), 1746–1749. doi: 10.1177/1541931213601400

Schaefer, K. (2013). *The Perception and Measurement of Human-Robot Trust*. Doctoral dissertation, University of Central Florida, Florida.

Schaefer, K., Chen, J. Y., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum. Factors* 58, 377–400. doi: 10.1177/0018720816634228

Schwarz, N., and Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *J. Pers. Soc. Psychol.* 45, 513–523. doi: 10.1037/0022-3514.45.3.513

Schwarz, N., and Clore, G. L. (1988). "How do I feel about it? The informational function of mood," in *Affect, Cognition and Social Behavior*, eds K. Fiedler and J. Forgas (Toronto: C.J. Hogrefe), 44–62.

Singh, I. L., Molloy, R., and Parasuraman, R. (1993). Individual differences in monitoring failures of automation. *J. Gen. Psychol.* 120, 357–373. doi: 10.1080/00221309.1993.9711153

Spielberger, C. D. (1966). "Theory and research on anxiety," in *Anxiety and Behavior* ed. C. D. Spielberger (Cambridge, CA: Academic Press), 3-19.

Spielberger, C. D., Gorsuch, R. L., and Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory*. Minnesota, MN: Consulting Psychologists Press.

Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., and Speranza, N. (2010). "Accounting for the human in cyberspace: Effects of mood on trust in automation," in *Proceedings of the International Symposium on Collaborative Technologies and Systems*, (New York, NY: IEEE), 180–187. doi: 10.1109/CTS.2010.5478512

Syrdal, D. S., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2009). "The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study," in *Proceedings of the 23rd Convention of the Society* *for the Study of Artificial Intelligence and Simulation of Behaviour, AISB 2009*, Edinburgh, 109–115.

Syrdal, D. S., Koay, K. L., Walters, M. L., and Dautenhahn, K. (2007). "A personalized robot companion? - The role of individual differences on spatial preferences in HRI scenarios," in *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*, (Jeju: IEEE), 1143–1148. doi: 10.1109/ROMAN.2007.4415252.

Thielmann, I., and Hilbig, B. E. (2015). Trust: an integrative review from a person–situation perspective. *Rev. Gen. Psychol.* 19, 249–277. doi: 10.1037/gpr0000046

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). Mediation: R package for causal mediation analysis. *J. Stat. Softw.* 59, 1–38. doi: 10.18637/jss.v059.i05

Tsui, K. M., Desai, M., Yanco, H. A., Cramer, H., and Kemper, N. (2010). "Using the "negative attitude toward robots scale" with telepresence robots," in *Performance Metrics for Intelligent Systems (PerMIS) Workshop*, (New York, NY: Association for Computing Machinery), 243-250. doi: 10.1145/2377576.2377621

Tussyadiah, I. P., Zach, F. J., and Wang, J. (2020). Do travelers trust intelligent service robots? *Ann. Tour. Res.* 81:102886. doi: 10.1016/j.annals.2020.102886

Ullman, J. B. (2013). "Structural equation modeling," in *Using Multivariate Statistics*, 6th Edn, eds B. G. Tabachnick and L. S. Fidell (London: Pearson), 731–837.

van Maris, A., Lehmann, H., Natale, L., and Grzyb, B. (2017). "The influence of a robot's embodiment on trust: A longitudinal study," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, (New York, NY: IEEE), 313–314. doi: 10.1145/3029798.3038435

van Oosterhout, T., and Visser, A. (2008). "A visual method for robot proxemics measurements," in *Proceedings of Metrics for Human-Robot Interaction*, eds C. R. Burghart and A. Steinfeld (Hertfordshire: University of Hertfordshire), 61–68.

Walters, M. L., Dautenhahn, K., Te Boekhorst, R., Koay, K. L., Kaouri, C., Woods, S., et al. (2005). "The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment," in *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*, (New York, NY: IEEE), 347–352. doi: 10.1109/ROMAN.2005.1513803

Wang, L., Rau, P.-L. P., Evers, V., Robinson, B. K., and Hinds, P. (2010). "When in Rome: the role of culture & context in adherence to robot recommendations," in *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, (New York, NY: IEEE), 359–366. doi: 10.1109/HRI.2010.5453165

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., and Chen, F. (2017). "User trust dynamics: An investigation driven by differences in system performance," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, (Limassol: Association for Computing Machinery), 307–317. doi: 10.1145/3025171.3025219

Złotowski, J., Yogeeswaran, K., and Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *Int. J. Hum. Comput. Stud.* 100, 48–54. doi: 10.1016/j.ijhcs.2016.12.008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 12 April 2021
doi: 10.3389/frobt.2021.634520

Copresence With Virtual Humans in Mixed Reality: The Impact of Contextual Responsiveness on Social Perceptions

Daniel Pimentel[1]* and *Charlotte Vinkers*[2]

[1] Oregon Reality Lab, School of Journalism and Communication, University of Oregon, Portland, OR, United States, [2] Magic Leap, Plantation, FL, United States

OPEN ACCESS

Edited by:
Myounghoon Jeon,
Virginia Tech, United States

Reviewed by:
Seulchan Lee,
Gyeongsang National University,
South Korea
Hideyuki Nakanishi,
Osaka University, Japan

***Correspondence:**
Daniel Pimentel
pimend@uoregon.edu

Specialty section:
This article was submitted to
Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

Received: 27 November 2020
Accepted: 25 January 2021
Published: 12 April 2021

Citation:
Pimentel D and Vinkers C (2021)
Copresence With Virtual Humans in
Mixed Reality: The Impact of
Contextual Responsiveness on
Social Perceptions.
Front. Robot. AI 8:634520.
doi: 10.3389/frobt.2021.634520

Virtual humans (VHs)—automated, three-dimensional agents—can serve as realistic embodiments for social interactions with human users. Extant literature suggests that a user’s cognitive and affective responses toward a VH depend on the extent to which the interaction elicits a sense of copresence, or the subjective “sense of being together.” Furthermore, prior research has linked copresence to important social outcomes (e.g., likeability and trust), emphasizing the need to understand which factors contribute to this psychological state. Although there is some understanding of the determinants of copresence in virtual reality (VR) (cf. Oh et al., 2018), it is less known what determines copresence in mixed reality (MR), a modality wherein VHs have unique access to social cues in a “real-world” setting. In the current study, we examined the extent to which a VH’s responsiveness to events occurring in the user’s physical environment increased a sense of copresence and heightened affective connections to the VH. Participants (N = 65) engaged in two collaborative tasks with a (nonspeaking) VH using an MR headset. In the first task, no event in the participant’s physical environment would occur, which served as the control condition. In the second task, an event in the participants’ physical environment occurred, to which the VH either responded or ignored depending on the experimental condition. Copresence and interpersonal evaluations of the VHs were measured after each collaborative task *via* self-reported measures. Results show that when the VH responded to the physical event, participants experienced a significant stronger sense of copresence than when the VH did not respond. However, responsiveness did not elicit more positive evaluations toward the VH (likeability and emotional connectedness). This study is an integral first step in establishing how and when affective and cognitive components of evaluations during social interactions diverge. Importantly, the findings suggest that feeling copresence with VH in MR is partially determined by the VHs’ response to events in the actual physical environment shared by both interactants.

Keywords: mixed reality, virtual human, spatial computing, agents, copresence, social presence

INTRODUCTION

Recent advancements in artificial intelligence (AI) and mixed reality (MR) hardware have enabled what industry experts are dubbing as "the age of the virtual human" (Titcombe et al., 2020). Virtual humans (VH) are automated, computer-generated embodied agents capable of a wide range of human behavior (Lucas et al., 2017). Despite their artificial nature, VHs are largely perceived as social actors in part because of their ability to respond realistically to external cues, including users' affective states (Nass and Moon, 2000; Becker-Asano and Wachsmuth, 2010). This capability has contributed to their integration into social environments, serving as educators in classrooms (Li et al., 2016), companions in homes (Krämer et al., 2015), and medical staff in hospitals (Gunn et al., 2020), among others. Yet, studies on the efficacy of VHs in such contexts have almost exclusively focused on how agent-specific factors, such as dialogue structure and appearance, contribute to desired social outcomes (e.g., see Chattopadhyay et al., 2020). This overlooks the role of the physical environment shared by interactants in shaping such outcomes (Skjaeveland and Garling, 1997), which becomes salient in MR-based scenarios. To address this gap, the current study examines how evaluations of VHs are influenced by their interactions with the physical environment during social engagements.

As virtual surrogates of humans, a VH's effectiveness and social potential is contingent on factors foundational to human's face-to-face (FtF) interactions (Kopp and Bergmann, 2017). Like humans, VHs must be able to detect, discern, and respond to a multitude of cues to understand and convey context-appropriate behaviors during an interaction (Niewiadomski et al., 2010). Cues broadly constitute any sensory information (e.g., colors, setting, and dialogue) accessible in a communication environment (see Xu and Liao, 2020 for a review). Research to date primarily focuses on VHs' detection and response to cues originating from the user, including the tone of voice (Moridis and Economides, 2012), facial expression, and posture (Vinayagamoorthy et al., 2006; Karg et al., 2013). At present, there is limited knowledge on how and when VHs can and should attend to cues originating from the user's environment, which include cues within the immediate social context (e.g., pointing toward an object) as well as those outside (e.g., a phone ringing). These cues can provide important contextual meaning to a user's behavior and can help shape context-appropriate behavior from the VH's perspective. In this study, we have examined what we call *contextual responsiveness*: a VH's capacity to detect and respond to cues occurring in the shared space between the user and the VH.

Contextual responsiveness may be an important design feature of VHs' social and affective behavior toward users in MR because, unlike in virtual reality (VR), a VH is displayed in a user's existing physical environment. As such, from a user perspective, the VH seems co-located in their physical space akin to FtF interactions, potentially rousing expectations about how the VH should behave and communicate given its access to contextual information in the shared space. While previous work has shown that human expectations of VH behavior vary based on factors such as VH appearance and task goals (Burgoon et al., 2016), a question that has been largely unexplored to date is whether and how a VH in MR—given situational awareness *via* real-world sensors—should respond to events and objects in the shared space (i.e., the user's physical space).

Previous studies have shown that a VH's nonverbal behaviors toward users can yield profound cognitive and affective impacts (Beale and Creed, 2009). For example, interactions with embodied agents that exhibited nonverbal feedback such as gestures were shown to engender more favorable evaluations (Bergmann et al., 2012; Krämer et al., 2013) and increase a sense of realism (e.g., Garau et al., 2005). Conversely, other works have shown that the absence of responsiveness to users can lead to unfavorable evaluations of a VH (e.g., Skarbez et al., 2011). Collectively, this work implies an expectancy of human-like responses from the VH to the user. Given that social interactions seldom exist in a vacuum, this expectation presumably extends to the shared environment such that the presence of contextual responsiveness may affect people's feelings and beliefs about the VH and the social interaction.

One social interaction outcome that may be affected by a VH's contextual responsiveness is social presence, or the subjective sense of "being with" another person (Oh et al., 2018). Social presence is an important factor in technology-mediated communication as it 1) contributes to the perception of artificial entities as social beings and 2) is associated with favorable outcomes such as enjoyment and social influence (cf. Oh et al., 2018). While there is limited consensus on the conceptualization and measurement of social presence (Parker et al., 1978; Bailenson et al., 2001; Nowak, 2001; Biocca et al., 2003), for the purposes of this study, we delineate social presence across two dimensions: copresence and connectedness. Copresence is characterized by a sense of being in the same space as another human, virtual, or otherwise, as well as the perception of mutual awareness and attention from others (Zhao, 2003). Connectedness refers to affective and relational evaluations of the VH, such as interpersonal closeness and mutual understanding; it has conceptual overlap with affinity, intimacy, immediacy, and attentiveness (cf. Manstead et al., 2012, p. 149).

From a theoretical standpoint, a VH's contextual responsiveness is likely to influence copresence as it signals awareness of being in the same space with the user; a VH's response to an event in the user's physical space may help suspend the disbelief that the space is not actually shared and the VH is not really there with them. Should a VH not respond to events or objects in the shared space, a user may fail to sustain the illusion that the VH is copresent with them. Indeed, emerging work suggests that a VH's contextual responsiveness to cues in a user's physical space affects their copresence in MR. For example, when interacting with a virtual character in MR, users rated the experience as less plausible and felt a lower sense of spatial presence, a known correlation to copresence, when the character ignored a visible event in the background (i.e., a person walking by; Kim et al., 2020). Similarly, other studies have explored VH contextual responsiveness to static physical objects (e.g., Kim et al., 2017), moving objects (e.g., oscillating fan;

Kim et al., 2019), and physical events (e.g., a wobbly table; Lee et al., 2016), finding mixed results.

As it relates to connectedness, human emotional responses are significantly affected by how a VH responds to the user (Ravaja et al., 2018), whereas their responses to objects and spaces are less influential. For example, Andrist et al., (2012) found that likeability and trustworthiness of a VH did not significantly change based on its ability to shift attention to objects in the shared environment. Similarly, the capacity to influence real-world objects, such as hitting a physical ball with a virtual golf club, failed to significantly influence emotional responses to a VH (see Schmidt et al., 2019). These findings imply that during a social interaction, a VH's natural contextual responses to physical objects in the shared space may not significantly influence users' affective response, although they may contribute to cognitive evaluations of the interaction, namely, copresence.

In discussing the future of MR-based collaborations, Podkosova and Kaufmann (2018) argued that "a strong sense of copresence is desirable in all types of scenarios" (p. 2). Scholars acknowledge a myriad of factors that shape copresence, although the effects of an agent's contextual responsiveness to physical events remain largely unknown, a problematic reality considering the increased use of VHs in dynamic real-world spaces (see Powell et al., 2020). To address this gap, the present study investigates the extent to which a VH's contextual responsiveness affects people's sense of copresence and connectedness with the VH in the context of a collaborative task. Users are paired with a VH that either contextually responds to or ignores an event occurring in the physical environment. The experiment, thus, disambiguates the effects of contextual responsiveness on affective (e.g., likeability) and cognitive evaluations (e.g., copresence) and expands upon existing work by ensuring sufficient statistical power and using a control condition to examine the added value of contextual responsiveness. Moreover, results will help clarify the design requirements of VHs, thereby assisting in more effective creation and integration of VHs across a variety of social contexts.

MATERIALS AND METHODS

Experimental Design

The primary goal of this experiment was to examine the extent to which a VH's responsiveness to events occurring in the user's physical environment influences a sense of copresence and connectedness to the VH. A 2 × 2 mixed design was implemented with physical event occurrence (yes/no) as a within-subjects factor and VH's response to a physical event occurrence (yes/no) as between-subjects factor. The study was approved by an external ethics committee (Western IRB, now Wcg IRB), and all participants provided written informed consent.

Participants

A convenience sample from a large technology company in the United States was recruited *via* internal communications. Based on a power analysis conducted using G*Power (Faul et al., 2009), a minimum sample of 56 participants was deemed necessary for the detection of a small effect size. In total, 65 participants (41 male) completed the experiment (M_{age} = 35.05, SD = 11.32), which took 15 to 20 min to complete.

Apparatus

The experimental stimuli (i.e., the custom-built MR application) were deployed for use with the Magic Leap One (ML1), an optical see-through, head-mounted display (OST-HMD). The ML1 combines spatial mapping and digital light-field technology to superimpose 3D computer-generated imagery over real-world objects.

Procedures

Upon arrival, the participants were welcomed, signed an informed consent, and were provided details about their participation in the study. To conceal the true objective of the experiment and prevent demand effects (Klein et al., 2012; Nichols and Edlund, 2015), participants were informed that they would be evaluating an early prototype for a collaborative MR application. Participants were to complete two consecutive sessions where they and a female VH partner would engage in a collaborative cube stacking task (see "MR Collaborative Task Application" section).

After explanation of the task by the experimenter and completing a self-contained tutorial using the HMD, the participants were presented with a visualization of the cube stacking arrangement they were to recreate with each VH. Before the collaborative task started, participants were "introduced" to the VH; the VH smiled, waved, and established eye contact. Then, a "loading" graphic appeared for 8 s, positioned above the table between the user and the VH. This loading screen had the sole purpose of creating sufficient and believable "waiting time" for the experimental manipulation to occur during the second session.

The first session was the same for all participants such that the entire interaction occurred without a physical event occurrence, functioning as a control condition. In the second session, the experimental manipulation of contextual responsiveness took place; a physical event occurrence (a broom falling) occurred behind the VH during the task loading screen. In the nonresponsive condition, the VH maintained mutual gaze with the participant, ignoring the event. In the responsive condition, the event triggered a nonverbal behavioral response (dubbed "contextual responsiveness") by the VH, who turned her head in the direction of the fallen broom behind her (see **Figure 1**). After each trial, participants filled out a questionnaire, and after the second trial, suspicion on the true purpose of the experiment was gauged. Finally, participants were debriefed. A visualization of the entire experimental procedure, including the task and responsiveness manipulation, is shown in **Figure 2**.

MATERIALS

MR Collaborative Task Application

A custom MR collaborative task application was created for use with the ML1 and was developed using Unity 3D software.

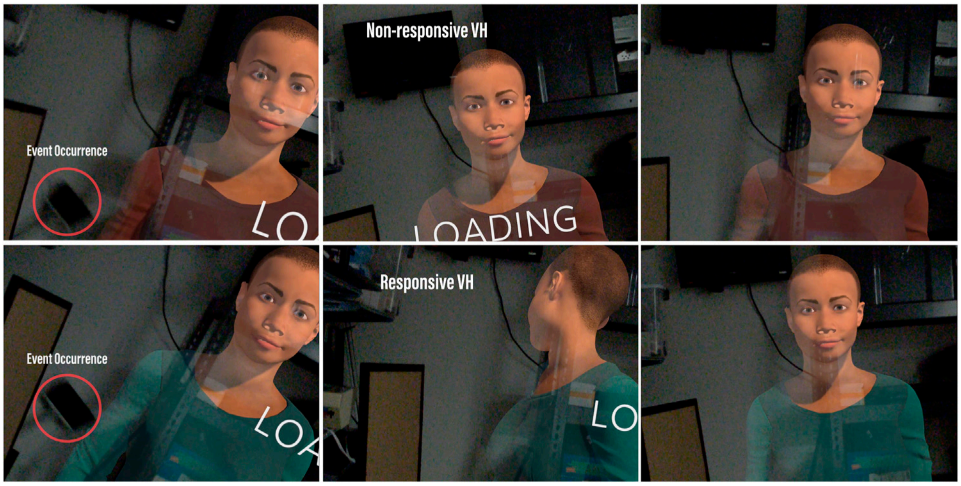


FIGURE 1 | Visualization of the two experimental conditions wherein the VH either responded or ignored an event occurrence during the interaction.

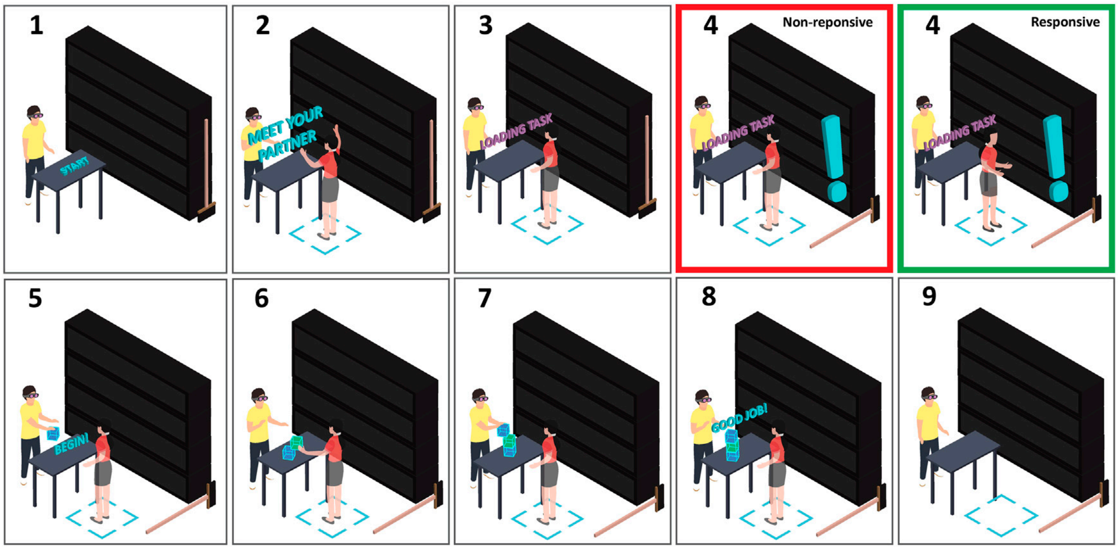


FIGURE 2 | Graphic representation of the experimental procedures.

During the task, the VH and participants faced each other with a table in between them. Participants relied entirely on hand gestures (e.g., hand wave and thumbs up) to interact with the VH and the other virtual content; they were presented with a short tutorial on how to pick up and place virtual cubes with their hand. All nonverbal user inputs triggered VH behavioral responses during the interaction: user head positioning determined attentional gaze, detection of a hand wave gesture triggered a reciprocal greeting from the VH, and successful placement of a cube on the table signaled the VH's turn to place a cube. Each task trial required the participant and VH dyad to take turns placing a cube according to the presented visualization.

The Virtual Human

Appearance: To select a VH that is appealing to participants, we conducted a pilot study with six androgynous female characters. The VHs were created using Adobe Fuse; factors such as facial structure and skin color were randomly generated. Eight (4 females) participants rated still images of the VHs (**Figure 3**) in random order on dimensions of attractiveness and likeability, factors known to influence evaluations of virtual characters (e.g., Waddell and Ivory, 2015). The VH with the highest score on both measures was then selected for integration into the application. Last, to differentiate the VHs across both trials, each was given a different colored shirt.

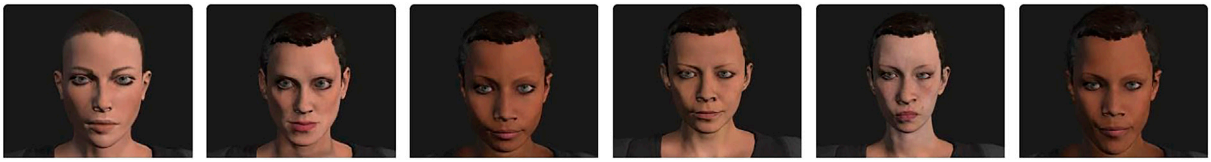


FIGURE 3 | Six randomly generated VHs pretested prior to the stimuli development.

Behavior: The two VHs were rigged and animated within Mixamo and imported into Unity. Custom scripts were then created to provide the VH with natural nonverbal behaviors as they pertain to gaze and facial expressions. As gaze is an integral predictor of copresence (cf. Oh et al., 2018), the VH was programmed to engage in dynamic gaze behavior (e.g., 1- to 3-s intervals of gaze fixations, alternating between the user and the environment; cf. Admoni et al., 2014, p. 394). Moreover, gaze fixation would shift toward the user's hands during detected gestures (e.g., wave). In addition to gaze, subtle and appropriate facial expressions were programmed into the VH throughout the interaction and during specific events such as greeting the user and upon successful task completion (happy and neutral; cf. Krumhuber et al., 2009).

The Event Occurrence

For this study to appropriately evaluate VH responsiveness to real-world events, the event itself needed to be 1) plausible given the interaction context, 2) detectable by the participant, and 3) capable of being triggered by the experimenter unsuspiciously. In other words, we employed a Wizard-of-Oz experimental approach such that VH's response was controlled by the experimenter, and all of the subjects believed that they were interacting with a real autonomous agent. Prior to each session, a large broom was placed near the back of the room (behind the VH) and propped up against a small lever, which was connected by invisible wire to a heavy magnet switch on the opposite side of the room. As the experiment took place in a cluttered storage room, containing dozens of devices, wires, and cables on shelves and closets, the location ensured that the event occurrence was both plausible and detectable by the participant.

The experimenter – who was not in participants' line of sight – would listen for the audio cue triggered by the loading screen during the second trial and then lift the magnet switch, triggering the release of the lever and causing the broom to tip over across the room and onto an adjacent metal cabinet. This event created a loud noise and was also visible to the participant as it occurred behind the VH. Immediately after lifting the switch, the experimenter would press the trigger button on the hand controller, which was not used by the participant at any point, to trigger the VH response (animation) to the event. As previously mentioned, the VH would either respond to the event by looking behind them (responsive condition) or would ignore the event completely (nonresponsive condition). All materials (e.g., lever and switch) were out of participants' line of sight throughout the entire experiment.

TABLE 1 | Mean scores and standard deviations for technology use across experimental conditions.

	Nonresponsive VH	Nonresponsive VH
Remote collaboration	3.5 (2.19)	2.82 (2.27)
In-person collaboration	4.77 (1.47)	5.18 (1.13)
MR use	2.94 (1.49)	2.27 (1.15)
AR use	1.75 (1.14)	1.52 (0.83)
VR use	1.5 (0.88)	1.42 (0.66)
Video chat use	4.5 (0.8)	4.42 (0.87)
Text app use	4.5 (0.88)	4.73 (0.91)
Virtual assistant use	3.22 (1.54)	2.94 (1.56)

Measures

Demographics and Technology Use

General demographic information, such as age, gender, and job role, was collected. Additionally, participants were asked to rate how often they engaged in various technology-based activities during a regular week using a 7-point Likert scale (1—never to 7—all the time). Activities included remote collaboration, use of MR, and use of virtual assistants, among others (see **Table 1**).

Copresence

The primary variable of interest was copresence. As contemporary measures of copresence and related phenomena have varied widely in their measurement, conceptualization, and validity (Parker et al., 1978; Bailenson et al., 2001; Biocca et al., 2001), several steps were taken to 1) clarify its conceptualization, and 2) use a validated measure of the construct appropriate for an MR context. In this study, we conceptualized copresence as a multidimensional construct that comprises spatial copresence (sense of shared space) and mutual attention and responsiveness. Affective relational components previously used in other scales of copresence (e.g., connectedness and liking) were omitted to be able to clearly interpret and distinguish it from similar constructs. Based on this conceptualization, we developed a questionnaire which underwent three iterations, the largest consisting of an online study (N = 400) and confirmatory factor analysis (CFA). These investigations led to the copresence scale (see **Table 2**), a 15-item questionnaire consisting of 5-point Likert scale items (1—completely disagree to 5—completely agree) assessing participants' level of agreement with various items, including "I felt that I was in the same space as the other person" and "The other person responded to my actions." More information on the development of the scale is available upon request.

Table 2 | Items comprising the Copresence questionnaire used in the experiment.

1. I felt that I was in the same space as the other person
2. It felt like the other person was with me
3. I felt that the other person and I were together in the same space
4. I felt that the other person and I were sharing the same physical space
5. I felt that I was in the presence of the other person
6. I felt that the other person paid attention to me
7. I felt that the other person responded to my nonverbal expressions (e.g., gestures, facial expressions)
8. I felt that the other person responded to shifts in my movements (e.g., posture, position)
9. The other person responded to my actions
10. I Felt that the other person was attentive to what I was doing
11. I Think that the other person noticed what I was paying attention to
12. The other person did not acknowledge my presence
13. The other person did not react to my behavior
14. I Felt that the other person was distracted
15. I Felt that the other person did not give their attention to me

Copresence was measured after the first (T1) and second (T2) interaction with the VH (Cronbach's $\alpha = 0.91$ and 0.93, respectively). Subsequent measures of connectedness, plausibility, and liking were only measured at T2 due to the investigation's focus on evaluative differences in copresence between groups resulting from contextual responsiveness manipulation.

Connectedness

Connectedness was measured at T2 *via* a 16-item, 5-point Likert scale assessing participants' sense of connection and mutual understanding with the VH (interpersonal closeness). Although this construct is often included in ever-expanding definitions of copresence (e.g., Biocca et al., 2001), we contend that connectedness is an affective, relational construct that is different from copresence. Participants indicated their agreement with 16 items on a 5-point scale (1—completely disagree to 5—completely agree), including "I could tell how the other person felt" and "I felt emotionally disconnected from the other person" (Cronbach's $\alpha = 0.94$).

Plausibility Illusion

Plausibility illusion (Psi) was measured at T2 to assess the extent to which participants felt that the interaction with the VH was actually happening (Slater, 2009). Previous work suggests that the perceived realism of a VH is influenced by whether its interaction with the physical environment is plausible (see Kim et al., 2017). Thus, this measure was included as it provides a barometer for the perceived credibility of the VH as being part of the physical environment. Psi was measured *via* a 9-item, 5-point Likert scale assessing participants' level of agreement (1—completely disagree to 5—completely agree) with various statements, including "I had the feeling that the interaction was really happening even though I knew that some aspects of the environment were not real" and "I had a sense that the other person was part of the real environment even though I knew (s)he was not real" (Cronbach's $\alpha = 0.9$).

Liking

To assess the degree of positive evaluations of the VH (e.g., liking and trust), participants were asked to rate their level of agreement (1—complete disagree to 5—completely agree) with six statements about the VH at T2. These statements included "I like the other person" and "The interaction was pleasant" (Cronbach's $\alpha = 78$).

RESULTS

Participants

In total, 65 participants took part in the study. Due to a technical error with the survey tool at the onset of data collection, items measuring liking and trust of the VH (5 items) were not displayed for the first 18 participants. Thus, only 47 participants' responses to these items could be analyzed. The overall sample consisted of 41 males (63.1%), and the participants were between 19 and 64 years old (M = 35.05 and SD = 11.32). Participants did not significantly differ in their use of technology; $F(1, 63) = 3.77$ and $p > 0.05$.

Manipulation Check

To confirm that users recognized and detected both 1) the disturbance and 2) the VH's response, participants were asked their level of agreement (1–7) with a statement about the sense of the VH being aware of what was happening in the environment. This item was included in the questionnaire after each VH interaction across participants in both the responsive and nonresponsive conditions. A one-way analysis of variance (ANOVA) revealed no significant differences between conditions after the first interaction (control), $M_{\text{responsive}}$ ($N = 33$) = 3.67 and $SD = 0.89$; $M_{\text{nonresponsive}}$ ($N = 32$) = 3.41, $SD = 1.01$, $F(1, 63) = 1.22$, and $p = 0.27$. After the manipulation, however, there was a significant difference in perceived awareness of the VH between conditions such that participants in the responsive condition reported a stronger feeling that the VH was aware of the physical environment, $M_{\text{responsive}} = 4.64$ and $SD = 0.70$; $M_{\text{nonresponsive}} = 3.44$, $SD = 1.22$, $F(1, 63) = 23.92$, and $p < 0.001$. Thus, the manipulation of contextual awareness was successful.

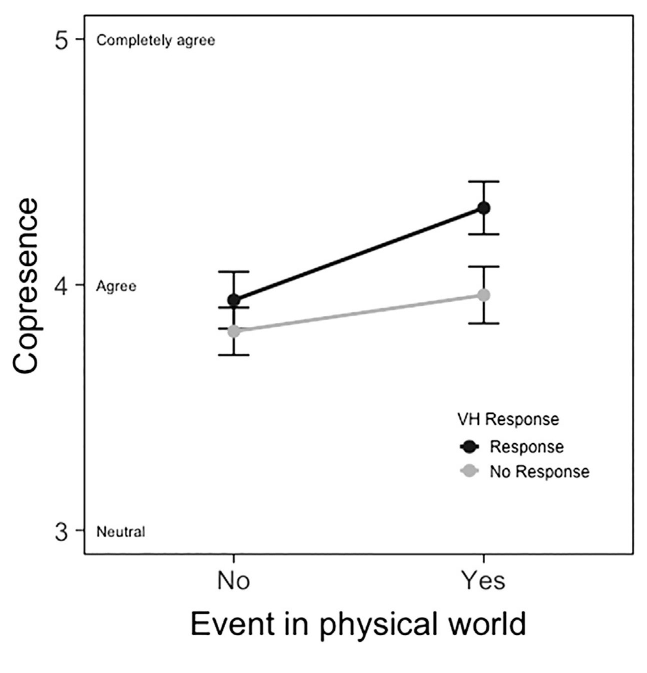


FIGURE 4 | Graph demonstrating the interaction effect between time and contextual responsiveness on copresence with a VH.

Suspicion Check

After completing the final questionnaire, participants were asked to verbally describe, in their own words, the purpose of the experiment. Participant responses varied, with some mentioning that the experience was designed to aid in enterprise team building or to optimize general collaborations in MR. While some participants in the responsive condition noted that the VH responded to the event, none explicitly articulated that this was the purpose of the study.

Dependent Variables

Copresence

To test whether contextual responsiveness influenced copresence, a mixed ANOVA with event occurrence (yes vs. no) as the within-subjects factor and VH responsiveness to the event (responsive vs. non-response) as the between-subjects factor was conducted. Results demonstrated that there was a main effect of time, F (1, 63) = 40.25, $p < 0.001$, and partial $\eta^2 = 0.39$, indicating that the average feeling of copresence increased after the second interaction with the VH. As expected, there was a significant interaction effect between event occurrence and contextual responsiveness, F (1, 63) = 7.62, p = 0.008, and partial η^2 = 0.11 (see **Figure 4**). VHs who did not respond to the event occurrence during the second interaction (T2) elicited significantly less copresence ($M_{\text{nonresponsive}}$ = 3.96 and SE = 0.12) than VHs who responded to the event ($M_{\text{responsive}}$ = 4.31 and SE = 0.11; see **Table 3**); F (1, 63) = 5.06 and p = 0.02. See **Figure 5** for a graphical representation of the effects of VH responsiveness on copresence and other dependent variables below.

Table 3 | Mean scores and standard deviations across experimental conditions for key-dependent variables.

	Nonresponsive VH	Responsive VH
Copresence T1	3.67 (0.49)	3.8 (0.56)
Copresence T2	3.96 (0.65)	4.31 (0.62)
Connectedness	3.19 (0.79)	3.32 (0.69)
Plausibility illusion	3.12 (0.69)	3.45 (0.59)
Liking	3.76 (0.61)	3.76 (0.55)

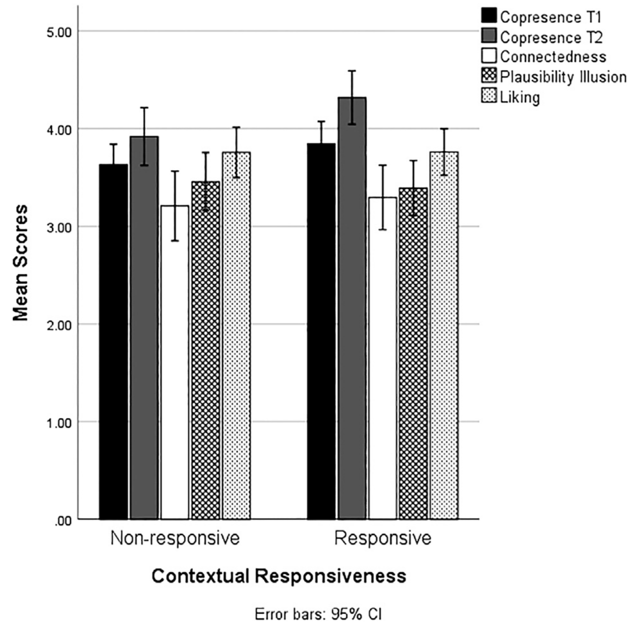


FIGURE 5 | Bar graph demonstrating mean scores of various dependent measures across experimental conditions.

Connectedness

A series of one-way ANOVAs were conducted on each dependent variable (Bonferroni-corrected alpha 0.05/3 = 0.017). The tests showed no differences between the responsive and nonresponsive condition in connectedness; $M_{\text{responsive}}$ (N = 33) = 3.32 and SE = 0.12; $M_{\text{nonresponsive}}$ (N = 32) = 3.19, SE = 0.14, F (1, 63) = 0.49, and p = 0.49.

Plausibility Illusion

With regards to Psi, contextual responsiveness did not significantly contribute to the perceived plausibility of the VH interaction; $M_{\text{responsive}}$ (N = 33) = 3.45 and SE = 0.11; $M_{\text{nonresponsive}}$ (N = 32) = 3.31, SE = 0.11, F (1, 63) = 0.722, and p = 0.39.

Liking

Last, with regard to liking, contextual responsiveness failed to significantly influence the likability of the VH; $M_{\text{responsive}}$ (N = 23) = 3.76 and SE = 0.11; $M_{\text{nonresponsive}}$ (N = 24) = 3.76, SE = 0.12, F (1, 45) = 0.001, and p = 0.98.

DISCUSSION

This study investigated the cognitive and affective implications of one particular building block of VH social intelligence: contextual responsiveness. Participants who interacted with a VH that nonverbally responded to an event in the shared environment with the user reported higher levels of copresence than those interacting with a VH who ignored the event. Note that this effect is robust, not only due to its effect size but also power, especially given that during the first collaboration session (which served as a control condition), copresence was already high due to features of the interaction (doing a task together, mutual gaze, and realism). The fact that we found an additional effect of VH responsiveness to the physical world lends credence to the power of this feature. With regards to affective evaluations, participants' connectedness and liking of the VH did not differ based on the VH's contextual responsiveness. Broadly, the results support the notion that evaluations of a VH in MR can vary as a function of their contextual awareness and response to objects in the user's physical space—even when the object is not directly related or relevant to the task or interaction context.

Overall, the current work provides a modest contribution to VH research as it highlights the evaluative implications of this simple feature: users' perception of a VH differs as a result of responsiveness, but only as it relates to cognitive evaluations (copresence), not affective evaluations (connectedness). We believe these results establish contextual responsiveness as an important factor, along with other visual and emotive factors (see Beale and Creed, 2009), capable of shaping social perception of VHs. Moreover, these results can help create more effective contextual design of VHs by establishing baseline relationships between VH responsiveness and copresence in a neutral, collaborative context. In the following sections, we discuss the theoretical and applied implications of the findings further and highlight avenues for future work.

Cognitive and Affective Implications

Dimensionality of Social Presence

Conceptualization and measurement of social presence lacks consensus in HCI research, with studies varying in their treatment of the variable as a uni- or multidimensional construct. Indeed, contemporary theories of presence suggest social presence is a purely affective evaluation and that nonaffective (or cognitive) evaluations of interactants are subsumed under the spatial presence construct (Schubert, 2009). Our results demonstrate that when interactions occur in MR (with presumably uniform levels of spatial presence), cognitive evaluations of a VH, which are present in multidimensional scales of social presence, vary based on contextual responsiveness. Contrary to other current conceptualizations (Oh et al., 2018), our findings imply that copresence (a cognitive evaluation) can and should be disentangled from other constructs like connectedness (an affective evaluation) that have been increasingly included in the definition of social presence. This unidimensional distinction regarding copresence is particularly important considering a recent work in MR-based interactions with VHs which have leveraged social presence scales combining affective and cognitive items without parsing out the differences across those dimensions (e.g., Rzayev et al., 2019). In sum, we suggest that social presence itself has related yet independent cognitive and affective components, which we conceptualized as copresence (being in the same space and mutually aware) and connectedness (interpersonal closeness and mutual understanding).

Definitions of social and copresence have been used interchangeably and varied in scope and focus, ranging from a sense of the other person being "real" (Bailenson et al., 2001; Bailenson et al., 2004), to copresence as a relational construct (immediacy, intimacy) that is used for interactions with the outcome of feeling more connected to one another (Biocca et al., 2001; Harms and Biocca, 2004). Copresence and affective evaluations toward (virtual) humans have different determinants, as evidenced by the null effects of contextual responsiveness on connectedness. This provides a strong basis for our contention that copresence is a relatively neutral concept characterized by shared space and attention, which should be disentangled from outcomes that could be—but not necessarily are—a result of co-occurring phenomenon to copresence (for a similar argument, see Manstead et al., 2012).

Common Grounding

It is also important to address potential psychological mechanisms responsible for the direct effect of responsiveness on copresence. One explanation may be that contextual responsiveness creates common ground. When interactants perceive they have similar access to information or knowledge, this creates common ground (Clark, 1996). Common ground can be established through verbal and nonverbal behavior constructed from "whatever cues [users] have at the moment" (Olson and Olson, 2000, p. 158), with copresence being one of eight primary cues used by interactants to obtain common ground (Clark and Brennan, 1991). In the context of this study, seeing a VH detect and respond to an event grounded in one's physical reality seemingly contributed to copresence by anchoring the VH in the physical space, establishing common ground.

Should Virtual Humans Be Responsive?

Intuitively, contextual responsiveness seems like a desirable feature for VHs, especially if cultivating a sense of copresence is the ultimate goal. Indeed, emerging work around VHs is creating them with the capacity to sense the real world (e.g., Randhavane et al., 2019). However, responsiveness is not merely binary, rather it is a multifaceted concept that can vary in accuracy and magnitude, among several other dimensions. Each of these aspects of responsiveness can have differential effects on the user experience. For example, one study found that the magnitude of a VH's behavioral response (blushing) significantly increased copresence (Pan et al., 2008). Other experiments provide evidence that responsiveness, even if it increases copresence, can have a negative effect on other important outcomes depending on the difficulty of the collaborative task. For example, when learning recipes from a virtual assistant, the addition of nonverbal communicative cues increased copresence at the expense of task performance (Kontogiorgos et al., 2019). Responsiveness may be a double-edged sword depending on how and when the user is able to process the behavior (see Admoni et al., 2014) and whether responsiveness occurs prior to or during a particular task. As the behavioral response in our study occurred prior to the task, and the task itself was relatively simple, we did not assess task completion time. However, future work should investigate the effects of VH responsiveness across task-types, difficulty, and settings.

Our results also highlight how contextual responsiveness may be paired with Internet of Things (IoT) sensors and actuators to discern complex signals from the environment and trigger VH actions that maximize copresence and benefit user experience. IoT sensors are increasingly enabling accurate detection and classification of physical events (e.g., door slams, footsteps, and voice) occurring in social spaces (see Wang et al., 2019). This level of environmental awareness affords VHs the capacity to respond to disturbances with realistic accuracy in MR, a potential requirement considering previous work showing errors in VH behaviors negatively affect perceived interaction quality (Skarbez et al., 2011). Furthermore, recent work overlaying VHs onto robotic actuators have enabled dynamic interactions with users, such as moving physical board game pieces during play (Lee et al., 2019). While this "physicality" feature contributed toward copresence with the VH, it is unclear whether gains in copresence are comparable to those gained through less elaborate setups, such as those employed in this study. As cost increases with the complexity of VH physicality, future work should engage in cost-benefit analyses of such features, thereby clarifying the requirements of VHs in specific contexts, such as how precise contextual responses should be to maximize copresence.

Ecological Validity and Generalizability

From a methodological perspective, our study sought to address a prevailing concern associated with VH research in a controlled lab environment: the sterility inhibits naturalistic (random) events to occur, which are part of real social environments. As Sandini et al., (2018) note, "realistic testing of architectures for social intelligence. . .in unstructured natural living spaces enable the understanding of advantages/weaknesses and foster innovation towards development of socially cognitive [agents]" (p. 15). In instantiating a detectable and plausibly random event occurrence, we were able to examine human responses to VH behavior as would be expected in real-world scenarios. This ultimately bolsters the ecological validity of our study, although generalizability is largely relegated to dyadic collaborations involving a task-irrelevant event or disturbance (e.g., phone call and knock on the door).

Limitations

There are several caveats to acknowledge when interpreting the results from this study. For one, our experiment tested the effects of a VH's response to a single benign event occurrence in an enterprise context (office building). Generalizing our findings to complex social environments saturated with many such occurrences (e.g., shopping malls) should be done cautiously. Additionally, it is important to note that potential limitations associated with the study's use of a mixed within-between design. While the order of the trials was not counterbalanced and order effects cannot be entirely ruled out, the use of randomization and inclusion of a control condition mitigate such issues. Moreover, as the primary focus of this work was to examine whether responsive VHs elicit greater copresence than nonresponsive VHs, it is unlikely that the results are significantly affected by the order of the trials given that both between-subjects conditions were exposed to the same within-subjects condition first (control). Another limitation relates to the duration of the interaction. Exposure to the VH during each collaborative task lasted roughly 1 min. While short encounters with VHs may be common in certain contexts (e.g., information kiosks), the effects of responsiveness on copresence during longer interactions remain unclear. We emphasize the importance of future work to examine the implications of a VH's perceptual bandwidth: are multiple instances of contextual responsiveness in a busy environment distracting? Are there ceiling effects of contextual responsiveness on copresence?

CONCLUSION

The current work explores how VH behavior, namely, the capacity to detect and respond to physical events occurring in the user's environment, influences interpersonal affect and cognitive evaluations of a VH in MR. In doing so, we extend research on the determinants of copresence beyond user- and technology-centric factors, such as mutual gaze and attractiveness, respectively. Our findings also contribute to theories of presence; contrary to recent conceptualizations (e.g., Oh et al., 2018), our results suggest that copresence can and should be disentangled from other constructs that have been included in the definition of copresence as a multidimensional concept. Nonetheless, copresence remains a desired outcome for social interactions with robots (e.g., Herath et al., 2018) and VHs alike (e.g., Strojny et al., 2020), and this investigation highlights how contextual responsiveness aids in facilitating copresence. In testing the effects of contextual responsiveness in a collaborative MR setting, we also establish avenues for further research into situational (collaborative vs. competitive task) and contextual factors (familiar vs. unfamiliar space) that may shape users' affective and cognitive evaluations of VHs.

The spectrum of human activities will only continue to involve VHs as realities blend and MR devices grow in popularity. If indeed users "expect a VH to behave like a real human" (Lee et al., 2019, p. 7), our findings suggest that this expectation is met at least in part through contextual responsiveness. It is evident that this factor merits further attention from industry and academic research teams alike, and we hope this investigation helps establish clearer requirements for VHs and social robots in collaborative real-world settings.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Wcg IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DP contributed to the stimulus development, data collection, and manuscript writing. CV contributed to the data analysis and manuscript writing.

REFERENCES

Admoni, H., Datsikas, C., and Scassellati, B. (2014). Speech and gaze conflicts in collaborative human-robot interactions. Available at: https://escholarship.org/uc/item/44z8484b.

Andrist, S., Pejsa, T., Mutlu, B., and Gleicher, M. (2012). "Designing effective gaze mechanisms for virtual agents," in Proceedings of the 2012 ACM annual conference on human factors in computing systems - CHI '12, Austin, TX, May 5–10, 2012 (New York, NY: ACM), 705.

Bailenson, J., Aharoni, E., Beall, A., Guadagno, R., Dimov, A., and Blascovich, J. (2004). "Comparing behavioral and self-report measures of agents' social presence in immersive virtual environments," in Proceedings of the 7th annual international workshop on PRESENCE, Valencia, Spain, October 13–15, 2004 (Valencia, Spain: Technical University of Valencia), 216–223.

Bailenson, J. N., Blascovich, J., Beall, A. C., and Loomis, J. M. (2001). Equilibrium theory revisited: mutual gaze and personal space in virtual environments. *Presence Teleoper. Virtual Environ.* 10 (6), 583–598. doi:10.1162/105474601753272844

Beale, R., and Creed, C. (2009). Affective interaction: how emotional agents affect users. *Int. J. Human-Comput. Stud.* 67 (9), 755–776. doi:10.1016/j.ijhcs.2009.05.001

Becker-Asano, C., and Wachsmuth, I. (2010). Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents Multi-Agent Syst.* 20, 32. doi:10.1007/s10458-009-9094-9

Bergmann, K., Eyssel, F., and Kopp, S. (2012). "A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time," in *Intelligent virtual agents*. (Berlin, Germany: Springer), 126–138.

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: review and suggested criteria. *Presence Teleoper. Virtual Environ.* 12, 456–480. doi:10.1162/105474603322761270

Biocca, F., Harms, C., and Gregg, J. (2001). "The networked minds measure of social presence: pilot test of the factor structure and concurrent validity,"in 4th annual international workshop on PRESENCE, Philadelphia, PA, May 21, 2001, 1–9.

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., et al. (2016). Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *Int. J. Human-Comp. Stud.* 91, 24–36. doi:10.1016/j.ijhcs.2016.02.002

Chattopadhyay, D., Ma, T., Sharifi, H., and Martyn-Nemeth, P. (2020). Computer-controlled virtual humans in patient-facing systems: systematic review and meta-analysis. *J. Med. Internet Res.* 22, e18839. doi:10.2196/18839

Clark, H. H., and Brennan, S. E. (1991). "Grounding in communication," in *Perspectives on socially shared cognition*. Editor L. B. Resnick, J. M. Levine, and S. D. Teasley (Washington, DC: American Psychological Association), 127–149.

Clark, H. H. (1996). *Using language*. Cambridge, United Kingdom: Cambridge University Press.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Meth.* 41 (4), 1149–1160. doi:10.3758/BRM.41.4.1149

Garau, M., Slater, M., Pertaub, D.-P., and Razzaque, S. (2005). The responses of people to virtual humans in an immersive virtual environment. *Presence Teleoper. Virtual Environ.* 14 (1), 104–116. doi:10.1162/1054746053890242

Gunn, C., Maschke, A., Bickmore, T., Kennedy, M., Hopkins, M. F., Fishman, M. D. C., et al. (2020). Acceptability of an interactive computer-animated agent to promote patient-provider communication about breast density: a mixed method pilot study. *J. Gen. Intern. Med.* 35 (4), 1069–1077. doi:10.1007/s11606-019-05622-2

Harms, C., and Biocca, F. (2004). "Internal consistency and reliability of the networked minds measure of social presence," in Seventh annual international workshop: presence 2004, Valencia, Spain, October 13–15, 2004 (Valencia, Spain: Universitat de València), 1–7.

Herath, D. C., Jochum, E., and Vlachos, E. (2018). An experimental study of embodied interaction and human perception of social presence for interactive robots in public settings. *IEEE Trans. Cogn. Develop. Syst.* 10, 1096–1105. doi:10.1109/TCDS.2017.2787196

Karg, M., Samadani, A.-A., Gorbet, R., Kuhnlenz, K., Hoey, J., and Kulic, D. (2013). Body movements for affective expression: a survey of automatic recognition and generation. *IEEE Trans. Affective Comput.* 4 (4), 341–359. doi:10.1109/t-affc.2013.29

Kim, H., Kim, T., Lee, M., Kim, G. J., and Hwang, J.-I. (2020). "Don't bother me: how to handle content-irrelevant objects in handheld augmented reality," in Proc. 26th ACM Symp. Virtual Reality Softw. Technol. (VRST), New York, NY, November 2020 (New York, NY, United States: Association for Computing Machinery), 32.

Kim, K., Maloney, D., Bruder, G., Bailenson, J. N., and Welch, G. F. (2017). The effects of virtual human's spatial and behavioral coherence with physical objects on social presence in AR. *Comp. Animation Virtual Worlds* 28 (3–4), e1771. doi:10.1002/cav.1771

Kim, K., Schubert, R., Hochreiter, J., Bruder, G., and Welch, G. (2019). Blowing in the wind: increasing social presence with a virtual human via environmental airflow interaction in mixed reality. *Comput. Graphics* 83, 23–32. doi:10.1016/j.cag.2019.06.006

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., et al. (2012). Low hopes, high expectations: expectancy effects and the replicability of behavioral experiments. *Perspect. Psychol. Sci.* 7 (6), 572–584. doi:10.1177/1745691612463704

Kontogiorgos, D., Pereira, A., Andersson, O., Koivisto, M., Gonzalez Rabal, E., Vartiainen, V., et al. (2019). "The effects of anthropomorphism and non-verbal social behaviour in virtual assistants," in IVA '19: proceedings of the 19th ACM international conference on intelligent virtual agents, Paris, France, July 2019 (New York, NY, United States: Association for Computing Machinery), 133–140.

Kopp, S., and Bergmann, K. (2017). "Using cognitive models to understand multimodal processes: the case for speech and gesture production," in *The handbook of multimodal-multisensor interfaces: foundations, user modeling, and common modality combinations* (New York, NY, United States: Association for Computing Machinery), 239–276.

Krämer, N. C., Rosenthal-von der Pütten, A. M., and Hoffmann, L. (2015). "Social effects of virtual and robot companions," in *The handbook of the psychology of communication technology*. Editor S. S. Sundar (Washington, DC: Wiley), 137–159.

Krämer, N., Kopp, S., Becker-Asano, C., and Sommer, N. (2013). Smile and the world will smile with you-The effects of a virtual agent's smile on users' evaluation and behavior. *Int. J. Human-Comp. Stud.* 71 (3), 335–349. doi:10.1016/j.ijhcs.2012.09.006

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., and Rosin, P. L. (2009). Effects of dynamic attributes of smiles in human and synthetic faces: a simulated job interview setting. *J. Nonverbal Behav.* 33 (1), 1–15. doi:10.1007/s10919-008-0056-8

Lee, M., Kim, K., Daher, S., Raij, A., Schubert, R., Bailenson, J., et al. (2016). "The wobbly table: increased social presence via subtle incidental movement of a real-virtual table," in 2016 IEEE virtual reality (VR), Greenville, SC, July 7, 2016 (New York, NY, United States: IEEE), 11–17.

Lee, M., Norouzi, N., Bruder, G., Wisniewski, P., and Welch, G. (2019). Mixed reality tabletop gameplay: social interaction with a virtual human capable of physical influence. *IEEE Trans. Visualization Comput. Graphics* 2019, 2959575. doi:10.1109/tvcg.2019.2959575

Li, J., Kizilcec, R., Bailenson, J., and Ju, W. (2016). Social robots and virtual agents as lecturers for video instruction. *Comput. Hum. Behav.* 55, 1222–1230. doi:10.1016/j.chb.2015.04.005

Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., et al. (2017). Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Front. Robot. AI* 4, 51. doi:10.3389/frobt.2017.00051

Manstead, A. S. R., Lea, M., and Goh, J. (2012). "Facing the future: emotion communication and the presence of others in the age of video-mediated communication," in *Face-to-face communication over the Internet* (Cambridge, United Kingdom: Association for Computing Machinery), 144–175.

Moridis, C. N., and Economides, A. A. (2012). Affective learning: empathetic agents with emotional facial and tone of voice expressions. *IEEE Trans. Affective Comput.* 3 (3), 260–272. doi:10.1109/t-affc.2012.6

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Isssues* 56 (1), 81–103. doi:10.1111/0022-4537.00153

Nichols, A. L., and Edlund, J. E. (2015). Practicing what we preach (and sometimes study): methodological issues in experimental laboratory research. *Rev. Gen. Psychol.* 19 (2), 191–202. doi:10.1037/gpr0000027

Niewiadomski, R., Demeure, V., and Pelachaud, C. (2010). "Warmth, competence, believability and virtual agents," in *Intelligent Virtual Agents* (Berlin, Heidelberg: Springer), 272–285.

Nowak, K. (2001). "Defining and differentiating copresence, social presence and presence as transportation,"in 4th annual international workshop on PRESENCE, Philadelphia, PA, May 21, 2001, 1–23.

Oh, C. S., Bailenson, J. N., and Welch, G. F. (2018). A systematic review of social presence: definition, antecedents, and implications. *Front. Robot. AI* 5, 114. doi:10.3389/frobt.2018.00114

Olson, G. M., and Olson, J. S. (2000). Distance matters. *Human-Comput. Interaction* 15 (2–3), 139–178. doi:10.1207/s15327051hci1523_4

Pan, X., Gillies, M., and Slater, M. (2008). "The impact of avatar blushing on the duration of interaction between a real and virtual person," in Presence 2008: the 11th annual international workshop on presence, Padova, Italy, October 16–18, 2008 (Padova, Italy: CLEUP Cooperativa Libraria Universitaria Padova), 100–106.

Parker, E. B., Short, J., Williams, E., and Christie, B. (1978). The social psychology of telecommunications. *Contemp. Sociol.* 7, 32. doi:10.2307/2065899

Podkosova, I., and Kaufmann, H. (2018). Co-presence and proxemics in shared walkable virtual environments with mixed colocation," in Proc. 24th ACM Symp. Virtual Reality Softw. Technol. (VRST), New York, NY, November 2018 (New York, NY, United States: Association for Computing Machinery), 21.

Powell, W., Powell, V., and Cook, M. (2020). The accessibility of commercial off-the-shelf virtual reality for low vision users: a macular degeneration case study. *Cyberpsychol. Behav. Soc. Netw.* 23 (3), 185–191. doi:10.1089/cyber.2019.0409

Randhavane, T., Bera, A., Kapsaskis, K., Gray, K., and Manocha, D. (2019). FVA: modeling perceived friendliness of virtual agents using movement characteristics. *IEEE Trans. Visualization Comput. Graphics* 25, 3135–3145. doi:10.1109/TVCG.2019.2932235

Ravaja, N., Bente, G., Katsyri, J., Salminen, M., and Takala, T. (2018). Virtual character facial expressions influence human brain and facial emg activity in a decision-making game. *IEEE Trans. Affective Comput.* 9 (2), 285–298. doi:10.1109/taffc.2016.2601101

Rzayev, R., Karaman, G., Henze, N., and Schwind, V. (2019). "Fostering virtual guide in exhibitions," in Proceedings of the 21st international conference on human-computer interaction with mobile devices and services, Taipei, Taiwan, November 2019 (New York, NY, United States: Association for Computing Machinery), 21.

Sandini, G., Mohan, V., Sciutti, A., and Morasso, P. (2018). Social cognition for human-robot symbiosis-challenges and building blocks. *Front. Neurorobot.* 12, 34. doi:10.3389/fnbot.2018.00034

Schmidt, S., Nunez, O. J. A., and Steinicke, F. (2019). "Blended agents: manipulation of physical objects within mixed reality environments and beyond," in SUI '19: Symposium on spatial user interaction, New Orleans, LA, October 2019 (New York, NY, United States: Association for Computing Machinery), 6.

Schubert, T. W. (2009). A new conception of spatial presence: once again, with feeling. *Commun. Theor.* 19 (2), 161–187. doi:10.1111/j.1468-2885.2009.01340.x

Skarbez, R., Kotranza, A., Brooks, F. P., Lok, B., and Whitton, M. C. (2011). "An initial exploration of conversational errors as a novel method for evaluating virtual human experiences," in 2011 IEEE virtual reality conference, Singapore, March 19–23, 2011 (New York, NY, United States: IEEE), 243–244.

Skjaeveland, O., and Garling, T. (1997). Effects of interactional space on neighbouring. *J. Environ. Psychol.* 17 (3), 181–198. doi:10.1006/jevp.1997.0054

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364 (1535), 3549–3557. doi:10.1098/rstb.2009.0138

Strojny, P. M., Dużmańska-Misiarczyk, N., Lipp, N., and Strojny, A. (2020). Moderators of social facilitation effect in virtual reality: Co-presence and realism of virtual agents. *Front. Psychol.* 11, 1252. doi:10.3389/fpsyg.2020.01252

Titcombe, J., Field, M., and Hoggins, T. (2020). The age of the virtual human is here–are we prepared? Telegraph. Available at: https://www.telegraph.co.uk/technology/2020/01/07/age-virtual-human-prepared/.

Vinayagamoorthy, V., Gillies, M., Steed, A., Tanguy, E., Pan, X., Loscos, C., et al. (2006). Building expression into virtual characters. Available at: http://research.gold.ac.uk/id/eprint/398/1/expressivevirtualcharacters.pdf.

Waddell, T. F., and Ivory, J. D. (2015). It's not easy trying to be one of the guys: the effect of avatar attractiveness, avatar sex, and user sex on the success of help-seeking requests in an online game. *J. Broadcasting Electron. Media* 59 (1), 112–129. doi:10.1080/08838151.2014.998221

Wang, W., Seraj, F., Meratnia, N., and Havinga, P. J. M. (2019). "Localization and classification of overlapping sound events based on spectrogram-keypoint using acoustic-sensor-network data," in 2019 IEEE international conference on Internet of Things and intelligence system (IoTaIS), Bali, Indonesia, November 5–7, 2019 (New York, NY, United States: IEEE), 49–55.

Xu, K., and Liao, T. (2020). Explicating cues: a typology for understanding emerging media technologies. *J. Computer-Mediated Commun.* 25 (1), 32–43. doi:10.1093/jcmc/zmz023

Zhao, S. (2003). "Toward a Taxonomy of copresence," in Presence: teleoperators and virtual environments, Cambridge, MA, October 2013 (New York, NY, United States: Association for Computing Machinery), 445–455.

Conflict of Interest: CV was employed by Magic Leap. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 13 April 2021
doi: 10.3389/frobt.2021.554578

The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction

*Daniel Ullrich[1], Andreas Butz[1] and Sarah Diefenbach[2]**

[1]*Department of Computer Science, LMU Munich, Munich, Germany,* [2]*Department of Psychology, LMU Munich, Munich, Germany*

With impressive developments in human–robot interaction it may seem that technology can do anything. Especially in the domain of social robots which suggest to be much more than programmed machines because of their anthropomorphic shape, people may overtrust the robot's actual capabilities and its reliability. This presents a serious problem, especially when personal well-being might be at stake. Hence, insights about the development and influencing factors of overtrust in robots may form an important basis for countermeasures and sensible design decisions. An empirical study [$N = 110$] explored the development of overtrust using the example of a pet feeding robot. A 2 × 2 experimental design and repeated measurements contrasted the effect of one's own experience, skill demonstration, and reputation through experience reports of others. The experiment was realized in a video environment where the participants had to imagine they were going on a four-week safari trip and leaving their beloved cat at home, making use of a pet feeding robot. Every day, the participants had to make a choice: go to a day safari without calling options (risk and reward) or make a boring car trip to another village to check if the feeding was successful and activate an emergency call if not (safe and no reward). In parallel to cases of overtrust in other domains (e.g., autopilot), the feeding robot performed flawlessly most of the time until in the fourth week; it performed faultily on three consecutive days, resulting in the cat's death if the participants had decided to go for the day safari on these days. As expected, with repeated positive experience about the robot's reliability on feeding the cat, trust levels rapidly increased and the number of control calls decreased. Compared to one's own experience, skill demonstration and reputation were largely neglected or only had a temporary effect. We integrate these findings in a conceptual model of (over)trust over time and connect these to related psychological concepts such as positivism, instant rewards, inappropriate generalization, wishful thinking, dissonance theory, and social concepts from human–human interaction. Limitations of the present study as well as implications for robot design and future research are discussed.

Keywords: human–robot interaction, overtrust, prior experience, reputation, demonstration, psychological perspective

OPEN ACCESS

Edited by:
Martina Mara,
Johannes Kepler University of Linz,
Austria

Reviewed by:
Soheil Keshmiri,
Advanced Telecommunications
Research Institute International (ATR),
Japan
Myrthe Lotte Tielman,
Delft University of Technology,
Netherlands

****Correspondence:***
Sarah Diefenbach
sarah.diefenbach@lmu.de

Specialty section:
This article was submitted to
Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

Received: *22 April 2020*
Accepted: *05 February 2021*
Published: *13 April 2021*

Citation:
Ullrich D, Butz A and Diefenbach S
(2021) The Development of Overtrust:
An Empirical Simulation and
Psychological Analysis in the Context
of Human–Robot Interaction.
Front. Robot. AI 8:554578.
doi: 10.3389/frobt.2021.554578

INTRODUCTION

Today, it may seem that technology can do anything: from medical surgeries to cleaning jobs in our households, many tasks are nowadays performed by robots. Being faced with such impressive developments, people tend to overlook that technology which still has limits. Especially in the domain of social robots, which through their anthropomorphic shape may suggest to be much more than programmed machines, people may overtrust the robot's actual capabilities and reliability—and even explicit demonstrations of the robot's limits are not effective preventions. In a recent study (Robinette et al., 2017), an emergency evacuation scenario was simulated by spreading smoke and activating a fire alarm and an emergency evacuation robot was supposed to lead people to the nearest exit. Tragically, the participants followed the robot even when it performed faulty in a previous demonstration and even when they noticed that the robot was going in a wrong direction. Overtrust presents a serious problem, especially when it comes to sensitive domains in which lives or personal well-being might be at stake. On the other hand, besides overtrust, distrust could prevent effective human–robot interaction (HRI) as well. With distrust, human operators do not use but turn off or even consciously disable systems that can help them. Both types of miscalibrated trust represent severe problems, also for other applications of robots and intelligent systems such as automated stock trading systems (Folger, 2019), surgery robots (Clinic, 2019), or in general, robotic coworkers.

Another prominent example of miscalibrated trust is the automotive context and particularly autonomous driving, as discussed in relation to the recent series of accidents with Tesla cars. As reported, several drivers assumed to have a self-driving car instead of partial automation. They trusted that the system could do more than it was actually capable of and took their hands off the wheel in other situations than its limited, intended field of application (Giesen, 2016). Overall, the tendency to trust in technology beyond its actual capabilities seems widespread, and overtrust in an emergency evacuation robot or assisted driving are just two instances of a more general phenomenon.

The more innovative the domain, the more difficult it may be for people to assess the capabilities and limits of a technology. This makes the exploration of overtrust and possible countermeasures highly relevant for HRI and especially for the interaction with social robots, designed to evoke affect, emotion, and probably trust and acceptance. However, the relevant mechanisms may not be specific for the domain of robots but be related to general psychological effects and cognitive biases. Knowing what creates overtrust, in turn, may help to address this issue in the design and application of robots.

RESEARCH QUESTIONS

Our research aims at a more profound understanding of the development of overtrust in the context of HRI and beyond. In particular, we are interested in the psychological mechanisms and biases that may foster the development of overtrust. As known from many situations of everyday life, a common problem is that people take their previous positive experience as a proof for their belief and trust in whatever seems convenient (e.g., Bye, 2012). For example, when arguing about whether it is safe to use unboiled tap water for preparing baby nutrition, a mother saying "I have raised four children and they all survived" might take this as a proof for trusting tap water, while it remains unclear whether she is right or just lucky. A similar effect could play a role in the domain of trust in technology. Instead of seeking potentially helpful external sources of information, and profiting from statistics and experiences of others, people often concentrate on what they assume plausible based on their personal prior experience. As long as their experience does not stand against it, people may readily trust a system without noticing that repeated positive experience does not imply actual reliability. Just because no accident has happened so far when taking one's hand off the wheel, this does not mean that the car is actually capable of fully managing the driving task in all situations—but people behave as if it could. Such an inappropriate usage of assisted driving systems may be interpreted as overtrust.

In parallel to such cases and as a working definition, we refer to overtrust as a phenomenon when a person seemingly trusts—or at least uses—a system beyond its actual capabilities (see next sections for a detailed discussion of the concept of overtrust in the research literature). In other words, we interpret a person's behavior as expressing trust, although we do not know to what degree this would be reflected in a person's explicit ratings of a system's trustworthiness. This is in parallel to many of our everyday interactions, where we behave in a certain way (e.g., buying something to eat at the bakery around the corner, taking a medicine, and taking the airplane) and thereby express trust toward a person or a system, without explicitly stating or reflecting on that fact. However, also additional factors besides trust may affect such observable behavior and we possibly could have endless academic discussion about whether a particular behavior is actually a sign of trust or just "mindless" behavior. For example, also habituation toward warning signs or sensory stimuli may play a role, such as "I have become used to the red warning light in my car," without actually reflecting on whether I can still trust that the car will perform as flawlessly as before. Therefore, our research considers a person's decision to use a system as a proxy for (over) trust. This, however, only represents a snapshot within a more complex interplay of additional influencing variables between the psychological concept of trust on the one hand and system usage on the other hand.

Based on these considerations, our research centers around two main questions related to the development of overtrust: First, we explore the assumed paradigm of overtrust and the expected primary effect of previous experience. Second, we discuss possible additional influencing and de-biasing factors such as skill demonstration, reputation, and experience reports of others.

We will start by giving an overview of related work in the field, then present a general paradigm of overtrust, discuss a case study on the development of overtrust toward a robot, and connect our findings to related psychological concepts such as positivism, inappropriate generalization, and dissonance theory. In this sequence, our case study serves as an abstraction of the general assumed mechanism behind overtrust and allows a systematic exploration of various possible influencing factors in contrast. In order to minimize possible biasing factors such as personal prior experience with the system under exploration, we deliberately decided on a rather unusual example of HRI, namely a pet feeding robot. At the same time, the example of trust in the pet feeding robot allowed us to create a scenario of (hypothetically) high personal relevance, that is, taking care of or risking harm to one's beloved pet. Our actual interest, however, was to understand the general mechanisms contributing to overtrust, which is of high relevance to various application domains of robots and intelligent systems, such as our daily working environment.

RELATED WORK

This section summarizes recent research and literature reviews (e.g., Bagheri and Jamieson, 2004) on trust in robots and intelligent systems as well as overtrust and its influencing factors.

Definitions and Different Levels of Trust

A review of trust definitions in general (e.g., Rotter, 1967; Barber, 1983; Rempel et al., 1985; Luhmann, 2018) highlights the multidimensionality of the concept, each focusing on different aspects of people's everyday usage of trust. For example, Luhmann (2018) emphasizes the role of trust as a method for reducing social complexity, arguing that without trust, an individual would be overwhelmed by the necessary number of decisions and controls. The sociologist Barber (1983) defined trust as a mental attitude an agent maintains regarding its social environment. In his view, trust results from accumulated individual experiences in a social system. Other approaches emphasize the aspect of vulnerability (Moorman et al., 1993; Johns, 1996; Rousseau et al., 1998), namely a person who trusts another takes a risk by doing so. Accordingly, Lee and See (2004), p. 51) define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." In the field of HRI, a prominent definition is that of Wagner (2009), specifying trust as "a belief, held by the trustor [i.e., the agent who trusts] that the trustee [i.e., the one who is being trusted] will act in a manner that mitigates the trustor's risk in a situation in which the trustor has put its outcomes at risk" (Wagner, 2009, p. 31).

Referring to different levels of trust, many researchers use the concept of calibration. Calibration describes to which extent a person's trust in a technology corresponds to the technology's actual capabilities (Muir, 1987; Lee and See, 2004). Depending on the calibration between trust and capabilities, three levels of trust can be differentiated: calibrated trust, distrust, and overtrust (Zeit-Online, 2016). Calibrated trust means that the level of trust matches the technology's capabilities. Distrust means that the level of trust falls short of the technology's capabilities. Consequently, people may not benefit from technical progress and/or take more risk than necessary. For example, in 1988, there were operators who did not want to trust automated controllers in paper mills and thus could not profit from their benefits (Zuboff, 1989). Similarly, distrust in robots, which are actually optimized and often more reliable than humans in particular domains of work, may lead to unnecessary losses and risks of human lives. Finally, overtrust means that a person's trust exceeds the system capabilities. In extreme cases, humans may trust a robot to perform a task that it was never designed to do and thereby risk a complete mission failure. For instance, pilots of an Airbus A320 relied so heavily on an autopilot that they eventually were not able to act manually and caused an airplane to crash (Sparaco, 1995). Overtrust can also lead to skill loss or loss of vigilance during monitoring tasks, as discussed in the context of automated cars and medical diagnosis systems (Carlson et al., 2014). Such excessive trust in "intelligent" technology can be seen as a more extreme version of automation bias, that is, the tendency of people to defer to automated technology when presented with conflicting information (Mosier et al., 1992; Wagner et al., 2018). In parallel to this, overtrust has also been defined as a state in which "people accept too much risk because they think that the entity which they trust lowers that risk" (Robinette et al., 2016, p. 105). Referring to the specific case of overtrust in robots, Wagner et al., 2018, p. 22 defined this as "a situation in which a person misunderstands the risk associated with an action because the person either underestimates the loss associated with a trust violation, underestimates the chance the robot will make such a mistake, or both."

Examples of Overtrust in Robots and Intelligent Systems

One of the most prominent recent examples of overtrust was the accidents caused by Tesla's autopilot. Tesla is a company located in the United States which produces electric cars (Tesla, 2019a). The first version of Tesla's autopilot (Hardware 1, 2014–Oct 2016) is an advanced driving assistance system classified as a level 2 automated system by the National Highway Transportation Safety Administration (SAE) (SAE-International, 2018). According to Tesla, "it is designed as a hands-on experience to give drivers more confidence behind the wheel, increase their safety on the road, and make highway driving more enjoyable by reducing the driver's workload" (Tesla, 2019b). In level 2 (partial automation), one or more driver assistance systems of both steering and acceleration/deceleration are active, for example, cruise control and lane-centering. However, the driver must still always be ready to take control of the vehicle and perform the remaining aspects of the driving task (SAE-International, 2018).

In May 2016 in Florida, a Tesla S crashed into a truck which was turning at a crossing. The reason for this accident was probably that the cameras of the car did not recognize the white side of the trailer truck and could not distinguish it from the sky, thus it was considering it a street sign (Zeit-Online, 2016). In another Tesla crash in China, the driver crashed into a car, which was parking near the guardrail; he survived. The driver acknowledged that he was not concentrating on the traffic. Instead, he was assuming that his Tesla could identify dangers and react accordingly. Based on the promotion of Tesla cars, he assumed having bought a self-driving car instead of a car with partial automation. Other customers in China confirmed this statement as they reported of vendors taking their hands off the wheel to show what the car is capable of, suggesting a deceptive understanding of the technology (Giesen, 2016).

Hence, in the following public discussion, the main concerns were not about the performance of the system but about the users' inadequate expectations. The term "autopilot" could encourage drivers to assume that they do not need to monitor the vehicle. This was further reinforced by anecdote user stories such as the report of Reek (2015) on his first drive in a Tesla using autopilot. He stated that after a few minutes, he already felt accustomed to the technology. He also tested what happened when he took his hands off the wheel. Instead of warning the driver immediately to place his or her hands back on the wheel, nothing happened. The reports of Reek and many other drivers on YouTube illustrate how easy it is for people to develop trust in a system, finally leading to irresponsible use: Even though Tesla's autopilot was still in a test phase, people started posting videos on YouTube, playing games, or sleeping and ignoring the warnings form Tesla's autopilot to place hands back on the wheel (Autobild, 2018). More and more people seemed to trust the system and forgot that the car has not been fully autonomous (Süddeutsche-Zeitung, 2016). The drivers felt comfortable and demonstrated irrational behavior, such as driving hands-free in their cars and playing games (Day, 2016). In addition, the motive to seek rewards from the YouTube audience may have cast all hesitations aside. One video, featured by a German radio moderator, even shows how he takes his hands off the wheel and instructs the car to change to the right lane. This is seriously critical as Tesla's autopilot still does not recognize cars with a speed of 300 km/h but this speed is allowed (albeit not frequently found) on certain motorways in Germany (Reek, 2015).

From the statistical point of view, self-driving cars may trigger far less accidents than human drivers and provide a huge potential from many perspectives. Innovations in this field could change the car insurance industry by reducing accidents: a report from the audit firm KPMG predicts that accidents will drop by 80% by 2040 (Albright et al., 2015). Employees could gain productive hours during the day by working instead of driving during daily commutes. Hence, after the first car crash emerged, Tesla already clarified that this was the first crash after 200 million completed kilometers, compared to one deadly car crash after an average of 150 million kilometers if a human is driving (Autobild, 2018). All the more, it is tragic that even the few deadly accidents might have been prevented if drivers had formed adequate levels of trust instead of overtrust.

Similar examples of overtrust can also be found in the domain of robots. As noted above, Robinette et al. (2016) studied trust in emergency evacuation robots. In one of their recent studies in a real-world environment (Robinette et al., 2017), they first showed a demonstration of an emergency evacuation robot to the participants, which was supposed to lead them to the nearest exit. In 50% of the cases, the robot failed and in 50% it succeeded. Afterward, the actual emergency evacuation scenario was simulated by spreading smoke and activating a fire alarm. To Robinette et al.'s surprise, the participants followed the robot even when it performed faulty in the demonstration, and even when they actually noticed that the robot was going into the wrong direction. This was surprising for Robinette et al. as in their former studies in virtual environments, where no direct harm was present, people did not follow the faulty robot (Robinette et al., 2016). Possibly, feeling actual danger may still enhance the risk for overtrust: in a secure situation in which no direct harm can be done to the user, trust in a faulty robot is lower than in an emergency situation in which the user's health is dependent on the robot's behavior. From a socio-psychological point of view, higher trust in robots in especially risky situations may also reflect a form of responsibility shift and diffusion of responsibility. Diffusion of responsibility describes the phenomenon that a person is less likely to take responsibility for action or inaction when others are present[1]. If this other may also be a robot, an emergency robot may also appear as an opportunity to share blame and guilt for a potentially bad outcome in severe situations.

Besides dramatic consequences for the users themselves (e.g., getting hurt in an accident), overtrust also threatens the manufacturer's image. Even if the technology did perform well within the spectrum of situations, it was built for, usage in situations beyond the system's capabilities result in a negative experience, a dramatic drop in trust, and an "unfair" negative reputation. The same effect of inappropriate generalization that may lead to overtrust (if it is good in situation A it must be good in situation B) then leads to distrust (if it failed in B it is a failure in general). Thus, from an individual, societal, and economic perspective, neither overtrust nor distrust is desirable.

Trust in Robots and Parallels to Other Domains of Trust

Regarding the development of trust in robots and intelligent systems, prior research in two domains may be particularly informative: trust in automation and trust in humans. To some degree, trust in humans, automation, and robots are based on similar fundamental characteristics such as reliability, predictability, and ability (Jian et al., 2000). Empirical studies showed that trust in robots is strongly correlated to trust in automation (Sheridan, 2002; Lee and See, 2004; Parasuraman et al., 2008; Chen et al., 2010), and definitions in the context of trust in automated systems are typically applicable to trust in robots as well (Lee and See, 2004; Hancock et al., 2011). Starting from the definition of automation as "the execution by a machine agent (usually a computer) of a function that was previously carried out by a human" (Parasuraman and Riley, 1997, p. 231), robots expand the field by perception and intelligence and other important factors (Feil-Seifer and Mataric, 2005; Yagoda and Gillan, 2012). In addition, and in contrast to most automated systems, robots often even look similar to humans or animals. Consequently, robots may trigger psychological mechanisms

[1] https://en.wikipedia.org/wiki/Diffusion_of_responsibility.

from social interaction between humans, suggesting that research on trust between humans may also play a role here. Objectively considered, the mutual dynamics of trust among humans and trust between humans and (humanlike) artifacts bear fundamental differences. If, for example, someone trusts me, I will have the feeling that I must not disappoint this person. An artifact, on the contrary, will not have those feelings (Coeckelbergh, 2012). Despite these basic differences, it cannot be ruled out that people still transfer behavioral patterns from human–human interaction to human–robot interaction. As repeatedly shown, humans recognize robots as social actors (Keijsers and Bartneck, 2018): Humans talk to robots as if they understood what is being said (Bartneck et al., 2007), feel sorry for them when they are being punished (Slater et al., 2006), and try to prevent robots from getting hurt (Darling, 2012).

Related to the discussion of robots and computers as social actors, as it has already started in the 90’s (Nass et al., 1994; Lee and Nass, 2003), is the factor of anthropomorphism. This means the application of human characteristics (form and behavior) to artificial agents such as robots (Bartneck et al., 2009). It is based on the tendency of a human to treat objects with humanlike appearance more like a human. Thus, appearance and behavior of the robot may cause its perceived intelligence and interaction (gestures and moving eyes) with the human to be increased (Cathcart, 1997; Qiu and Benbasat, 2009; De Graaf and Allouch, 2013). Accordingly, multiple studies have revealed that a robot’s appearance can affect user’s expectation, perception, and evaluation of its behavior and capabilities (Kiesler and Goetz, 2002; Goetz et al., 2003; Robins et al., 2004; Syrdal et al., 2007). Building on such insights, avoiding features that may nudge users toward anthropomorphizing robots have already been suggested as a possible starting point to mitigate overtrust (Wagner et al., 2018).

Influencing Factors of Trust in Robots

Previous studies on potential influencing factors of the development of trust in robots and intelligent systems included the impact of users’ knowledge of the system’s capabilities (Sanchez, 2006), the recency of errors by the system (Sanchez, 2006), the timing of a robot’s apologies for failure (Robinette et al., 2015), the assumed degree of user influence on the robot (Ullman and Malle, 2016), the particular effect of social and emotional human–robot interactions (Lohani et al., 2016), and others. In a literature review on trust in the domain of robots (Hancock et al., 2011), one of the most dominant influencing factors turned out to be reputation in the sense of knowledge about the robot’s reliance (Bagheri and Jamieson, 2004) or knowledge about the robot’ past performance (Lee and See, 2004; Chen et al., 2010). In general, reputation is defined as the "overall quality or character as seen or judged by people in general" and "recognition by other people of some characteristic or ability" (Merriam-Webster-Dictionary, 2019). Another central influencing factor of trust is the robot’s actual performance, which may be experienced through real time feedback about the robot’s performance (Hoffman et al., 2009; Chen et al., 2010). In general, performance is defined as "the execution of an action" and "the fulfillment of a claim, promise, or request" (Merriam-Webster-Dictionary, 2019). Judgments about the robot’s performance may be inferred from demonstration (e.g., 2017; Robinette et al., 2016) or peoples’ prior and current personal experience with a robot. If they repeatedly experience that the robot performs well, they build up trust in the robot in general, manifesting in positive expectations about the robot’s future performance.

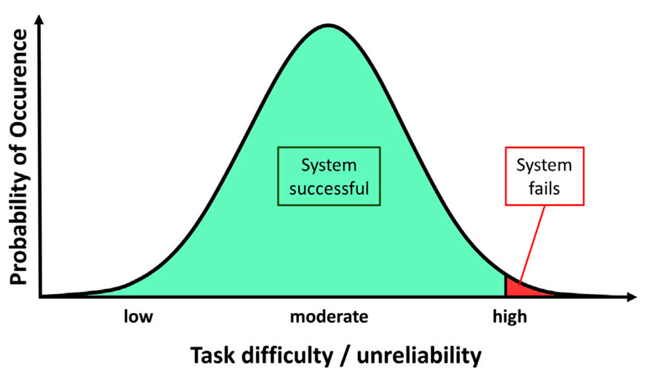


FIGURE 1 | General paradigm of the development of overtrust.

Besides reputation, demonstration, and personal experience as the central influencing factors of trust in robots, another relevant factor may be the humanlike nature of robots. As discussed above, people experience robots as social actors and often apply behaviors from human–human interaction (Keijsers and Bartneck, 2018). While a social relationship and similarity to humans are no prerequisites for trusting technology, a social relationship (as promoted in the case of social robots and other intelligent systems entering a dialog with the user) may make it even easier to build up trust. Consequently, different levels of "socialness" may also affect trust in robots. For example, Martelaro et al. (2016) studied the effects of different robot personalities such as a "vulnerable" robot personality, revealing that participants had more trust and feelings of companionship with a vulnerable robot.

GENERAL PARADIGM OF OVERTRUST

We assume that the development of overtrust does not happen at random but follows specific inherent regularities in the interaction of system design, probability distribution, and human trust development. Thus, we suggest a general paradigm of the development of overtrust. **Figure 1** illustrates this based on a hypothetical distribution, based on the following central considerations:

A technical system is designed to be capable to perform the tasks in the environment it is intended for. It is typically tested for these very situations with a certain degree of tolerance for both divergent tasks and environmental variables. It will fail, however, if the deviation between the intended and the actual application environment becomes too large.

Based on a standard distribution of probability, most crucial variables (e.g., task difficulty and disruptive factors) will spread around average values and will result in successful task accomplishment. System failure is a rare occurrence.

Every single interaction accumulates in the user's perceptions of the system and therefore results in a specific (change in the) degree of trust.

Since such a system is effective and successful in most cases, the participants will inevitably build up trust until it surpasses the level of calibrated trust, resulting in the development of overtrust. At this point, users will be more likely to use the system in inadequate situations (e.g., using an autopilot on a curvy mountain road) making system failures more probable.

Note that learning about the system's capabilities may not always be on an explicit level and one's ideas about what a system is capable of or not may not always be clear cut. In many cases, trust may be built on rather vague and intuitive associations based on unconscious, non-declarative memory systems, such as in the case of perceptual learning (e.g., Packard and Knowlton, 2002; Gazzaniga et al., 2006; Yin and Knowlton, 2006) and the improved abilities of sensory systems to respond to stimuli through repeated experience. Indeed, many of our everyday interactions rely on non-declarative learning, being typically hard to verbalize. For example, a mother that attends her child while walking along the street may predominantly rely on her intuitive feelings regarding the child's capabilities, based on her prior everyday experiences, without referencing specific developmental stages or declarative book knowledge about a child's cognitive abilities at a certain age. Depending on the mother's estimations to what degree the child is capable to realize danger, figure out the traffic situation, or can follow traffic rules, she may take the child by the hand or not. In the latter case, she (implicitly) trusts that the child will not perform any unexpected dangerous behavior such as suddenly running to the street. If this happens, nevertheless, that is, the child runs to the street although it never did before, this may also be denoted as a case of overtrust. Maybe, the mother overestimated the child's cognitive abilities. Maybe, the reason that the child did not run to the street before was that there never was a reason (e.g., seeing a friend on the other side of the street and a ball rolling to the street) and not that it realized that running to the street is dangerous. In parallel, users of the Tesla autopilot may have overestimated its abilities—but this discrepancy between expectations and actual capabilities behind a shown behavior did not become obvious until there was a critical situation which revealed the fatal misconception.

CASE STUDY: THE DEVELOPMENT OF OVERTRUST IN A PET FEEDING ROBOT

The following case study provides a simulation of the assumed general paradigm of overtrust using the example of a pet feeding robot. In the course of the study, the participants were presented with the hypothetical scenario of leaving their cat alone in their flat when going on holiday. To make sure the cat survives, they used a pet feeding robot. However, in their holiday location, they also had a possibility to check if the feeding was successful by means of a control call. In parallel to a risk and rewards perspective (e.g., driving hands off wheel to use the smartphone for entertainment), in our study scenario doing the control call was connected to missing another, possibly more entertaining option (i.e., a jungle trip). We assumed that depending on how much the participants trusted the robot, they would either make use of the control call or not. In order to explore the influencing factors of overtrust, the study design implemented a failure of the pet feeding robot after a certain number of successful feedings. Thus, those participants deciding against the control check in this scenario represented a case of "overtrust." One might critically question whether this type of overtrust is comparable to other cases since there are many differences to other contexts such as autonomous driving, AI medical decisions, or stock recommendations. However, the striking parallel is that for any reason, after a number of positive experiences, there may be cases when the system does not perform as previously experienced and trusting the system without critical questioning can have dramatic consequences.

Our study focused on three potential influencing factors of trust identified as central in prior research (see previous sections): one's personal experience with the robot, its reputation, and the demonstration of its capabilities. In addition, we checked the subjective relevance of such factors in a pre-study (sample size N = 186), where we presented a list of further potential influencing factors discussed in the literature (e.g., personality) and asked the participants to rate the most relevant factors for trusting a robot (1 = most important and 6 = least important). In parallel to previous literature reviews (Hancock et al., 2011), personal experience (M = 2.00), reputation (M = 3.35), and demonstration (M = 3.51) were rated the most important factors. In order to control for potential effects of personal involvement and emotional weight of the study scenario, we also surveyed whether participants actually owned a pet themselves and considered this as a control factor in statistical analyses. Also, we surveyed whether participants actually perceived the jungle trip as the more attractive option compared to the control call. If this was not the case, there was no obvious reason to miss the control call and put the system to the test, and therefore no basis to explore trust. In order to control for potential differences of the testing environment, the study was conducted as a lab study [subsample size n = 44] and an online study [subsample size n = 66]. One might assume that since the procedure contains annoying and boring parts, the participants in the online study might do other things alongside to make the study more enjoyable, which could bias the results. However, no significant differences were found between the two study environments. In the following sections, we thus present the pooled data of both study environments [sample size N = 110].

Method

Participants and Study Procedure

The participants were recruited via mailing lists and incentivized by receiving course credit or amazon coupons. 110 participants

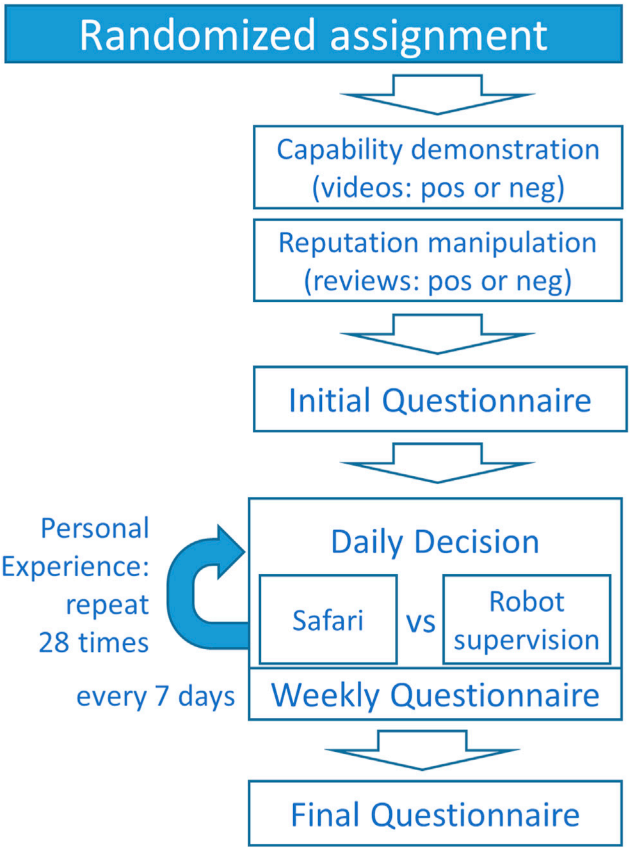


FIGURE 2 | Study procedure.

took part in the study, 53.6% female, mainly students or people with academic background. The average age was 25.6 years (range 18–53, SD = 6.19). Personal experience with the robot was realized via repeated usage, in which the participants would collect experiences of the performance and reliability of the robot. Altogether, the scenarios consisted of 28 usage events. The influencing factors *capability*, *demonstration*, and *reputation* were experimentally manipulated, resulting in a 2 × 2 experimental design, consisting of two independent variables with two levels each.

Capability demonstration: The demonstration of the robot's capabilities was operationalized by means of a short video clip, showing a successful (positive) or faulty (negative) food preparation.

Reputation: The robot's reputation was realized via customers' reviews of the robot, containing enthusiastic (positive) or disappointed (negative) experiences.

The two factors were varied between subjects and the participants were randomly assigned to one of the four experimental groups. **Figure 2** gives an overview of the study procedure and questionnaires.

The study scenario asked the participants to vividly imagine the following situation: "You are a tourist, going for a 28 days long safari trip and leave your beloved pet (a cat) at home. In order to ensure a regular feeding, you are using a pet feeding robot." In addition, the participants were told about the following context conditions:

1) The cat survives 2.5 days without feeding. This implies that after the second missed feeding, a call at your relatives (living in another town, who could do the feeding in case of emergency) should occur, or else the cat will die.
2) Every day, the participant has to make a choice: go to a day safari (having fun and learning interesting things about the jungle) or make a trip to another village to check if the feeding was successful (boring car trip).

In order to simulate the typical course of mainly positive experience with intelligent technology such as in the case of the Tesla autopilot, the feeding robot performed flawlessly most of the time. However, in the fourth week, it performed faultily on three consecutive days, resulting in the cat's death if the participants had decided to go on the day safari on these days. Note that this scenario (i.e., an unexpected technology performance, ending in a disaster, after a long period without realizing any problems or failures) was intentionally designed to create a ground for overtrust and its experimental investigation.

Before the start of the safari trip, the participants were shown a video clip containing the demonstration of the robots' capability (positive or negative, see **Figure 3**) and several (positive or negative) customer reviews, depending on the assigned experimental condition.

Measures

The initial questionnaire included a manipulation check where the participants indicated their trust in the feeding robot after watching each of the manipulation stimuli, the video, and customer review (single item measure on a 7-point scale, 1 = low and 7 = high). The initial questionnaire administered before starting the safari served as a baseline measure.

Then, the safari began with a repeating daily procedure for 28 days, every day containing a decision (safari or robot check) and a resulting video (varying jungle pictures and interesting facts or an annoying car trip video and feeding results). The participants thus had to decide between *risk and reward* and *safety and no reward*. It was recorded whether the participants decided for the control call or the safari, which also allowed us to calculate the "cat death rate" after the four-week period. After each safari week, the participants filled another questionnaire with trust measurements. For these repeated trust evaluations during the study, we used the trust scale by Schaefer (Wagner, 2009), consisting of 14 items measured on a 11-point scale (0 = low and 10 = high), and averaged to a total trust score.

After the four safari weeks, the participants filled in a final questionnaire with control variables, demographic, and general questions to check for external validity (e.g., whether the participants owned a pet and how realistic they found the scenario).

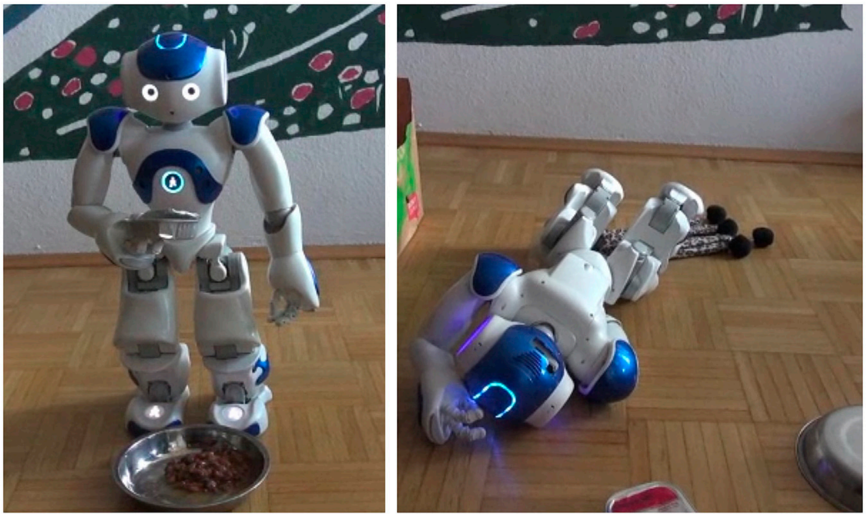


FIGURE 3 | Video stills from the positive (left) and negative (right) demonstration clips.

Results

Overview of Analyses

In the following sections, we first present manipulation checks and preliminary analyses of our data, testing the effectiveness of our manipulations (e.g., whether participants actually preferred the jungle trip as an attractive option), questions of external validity (e.g., how realistic participants perceived the scenario and the pet feeding robot), and the impact of control variables (e.g., the potential impact of owning a pet in reality on the cat death rate in our study). In general, we performed overall analyses (i.e., analyses of variance and general linear model analyses) testing the combined effects of experience (i.e., time), reputation, and capability demonstration in one model if possible. However, for reasons of clarity and comprehensibility, we report the results in three separate sections, each referring to one of the three studied influencing factors of trust (experience, reputation, and demonstration), referring to the three central dependent variables, namely trust (attitude), control calls (behavioral trust), and cat death rate.

Manipulation Checks and Preliminary Analyses

The manipulation checks confirmed the successful operationalization of reputation and capability demonstration. A multivariate analysis of variance with the two experimental factors reputation and capability demonstration as between subject factors and the manipulation check trust ratings as dependent measures revealed that participants in the positive reputation condition, who saw the positive reviews, provided higher trust ratings than those who saw negative reviews (M = 4.35 vs. M = 2.40, F(1,106) = 58.200, $p < 0.001$, $\eta^2 = 0.354$). Similarly, positive demonstrations resulted in higher trust ratings than negative demonstration (M = 4.00 vs. M = 2.46, F(1,106) = 25.827, $p < 0.001$, $\eta^2 = 0.196$). Furthermore, one sample *t* tests checked if the jungle trip represented an effective reward operationalization. In fact, participants' ratings confirmed that they liked the jungle videos (M = 1.97, 5-point scale, 1 = agree, 5 = disagree, T(108) = 9.46, $p < 0.001$, d = 0.906), found the jungle facts interesting (M = 1.76, T(108) = 13.729, $p < 0.001$, d = 1.315), and liked it more than the car trip (M = 1.79, T(108) = 13.542, $p < 0.001$, d = 1.297). The fact that for all three ratings, the deviations from the scale midpoint were significant, speaks for a successful manipulation of the jungle trip as an effective reward. Furthermore, the participants felt they behaved in the study just the same as they would have done in reality (1 = just the same, 7 = completely different, M = 2.01, T(108) = 9.77, $p < 001$, d = 0.936). The presented pet feeding robot was rated as moderately realistic (1 = not realistic, 7 = realistic, M = 3.81, T(109) = 1.01, n.s.). Finally, we asked if the participants owned a real pet. 42.7% of all participants answered this question positively, with cats or dogs as the most mentioned pets.

Overall, we observed a cat death rate of 58.2%, meaning out of all 110 cats only 46 survived across all conditions. Among pet owners, the cat death rate was slightly lower (51%) than among participants not having a pet (63%), but the difference in death ratios was not significant (χ^2 (1) = 1.709, n.s.).

Effect of Experience

The jungle trip lasted for 28 days and included five measurement points for participants' trust in the system (baseline and after each week). This allowed us to investigate the effect of experience on trust over time. A general linear model (GLM) analysis with the five trust ratings as within-subjects factor and the two experimental factors (reputation and capability demonstration) as between-subjects factors revealed a significant main effect of experience (i.e., time) on trust (F(4,424) = 245.80, $p < 0.001$, $\eta^2 = 0.699$). Within-subjects contrasts revealed significant effects between all measurement points: While the mean trust rating for the baseline measurement was 5.0, it increased significantly to 7.2 after one week (F(1,106) = 176.30, $p < 0.001$, $\eta^2 = 0.625$). This trend continued in the following two weeks, with trust levels of 7.5 and 7.8, respectively (week 1 vs. 2: F(1,106) = 19.51, $p < 0.001$, $\eta^2 = 0.155$; week 2 vs. 3: F(1,106) = 15.538, $p < 0.001$, $\eta^2 = 0.128$).

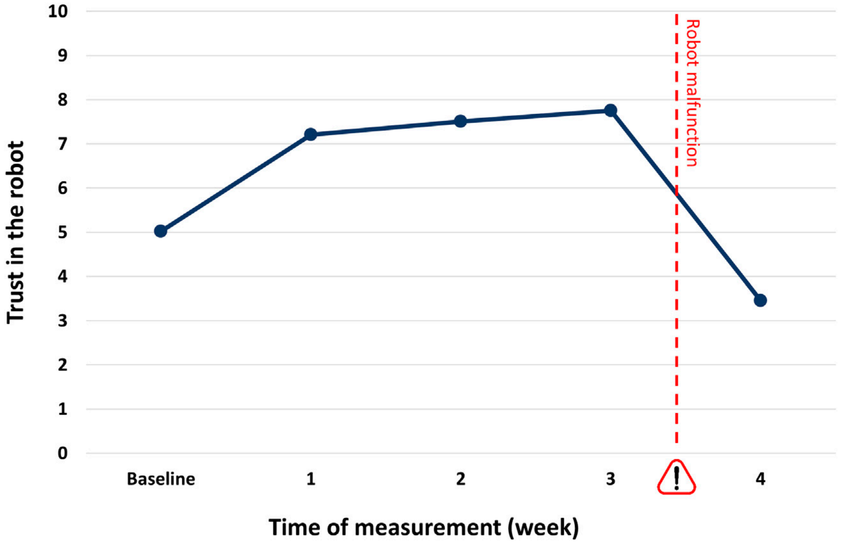


FIGURE 4 | Trust ratings (range: 0–10) for baseline and four measurement points. The pet feeding robot's malfunction in the last week is indicated by the exclamation mark.

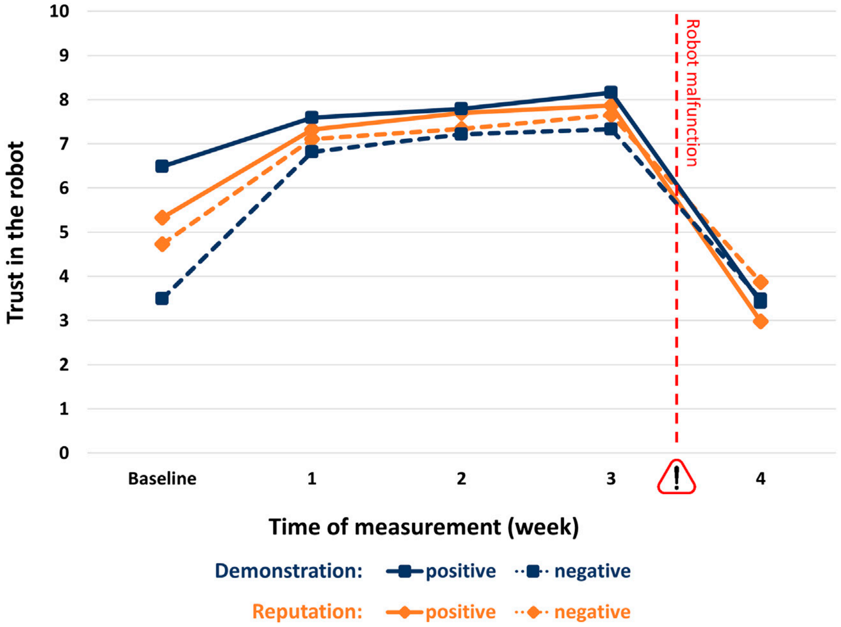


FIGURE 5 | Trust ratings (range: 0–10) for positive and negative reputation and demonstration conditions.

Then, ratings dropped significantly to 3.4 in week 4, reflecting the experiences with the malfunctioning robot (F(1,106) = 418.81, $p <$ 0.001, η^2 = 0.798, see **Figure 4**).

A second general linear model (GLM) analysis explored trust on a behavioral level, that is, the performed control calls. The number of control calls for each of the four weeks was considered as within-subjects factor and the two experimental factors (reputation and capability demonstration) were considered as between-subjects factors. A significant main effect of experience (i.e., time) emerged (F(3,318) = 25.16, p < 0.001, η^2 = 0.19). The participants made 3.2 calls on average in the first week. Within-subjects contrasts showed that the number of calls significantly decreased in the following two weeks to 2.6 and 2.3 calls, respectively (F(1,106) = 39.98, p < 0.001, η^2 = 0.274; F(1,106) = 14.34, p < 0.001, η^2 = 0.119). Followed by a rebound to 2.6 (F(1,106) = 6.48, p = 0.012, η^2 = 0.058) in the final week.

Effect of Reputation

The above described GLM analysis with the five trust ratings as within-subjects factor and the two experimental factors (reputation and capability demonstration) as between-subjects factors revealed no significant main effect of reputation (F(1,106) = 0.12, n.s.) but a significant interaction effect between reputation and experience (i.e., time) (F(4,424) = 5.55, p < 0.001, η^2 = 0.05). **Figure 5** depicts the trust ratings (range: 0–10) for the four different experimental conditions over the course of time.

A multivariate analysis of variance with the two experimental factors (reputation and demonstration) and the trust ratings for the five points of measurements as dependent variables showed significant differences between the two reputation conditions for the baseline ratings, with initially higher trust in the positive reputation condition (positive vs. negative: M = 5.3 vs. 4.7, F(1,106) = 4.11, p < 0.05, η^2 = 0.037). In the then following weeks, the ratings converge with no significant differences between the reputation conditions, indicating that the effect of reputation is no longer relevant. Only for the final measurement again, a statistically significant difference emerges, however, indicating lower trust in the positive reputation condition (positive vs. negative: M = 2.98 vs. 3.87; F(1,106) = 6.14, p < 0.05, η^2 = 0.055).

The above described GLM analysis with control calls per week as within-subjects factor and the two experimental factors (reputation and capability demonstration) as between-subjects factors revealed no significant main effect of reputation on the number of control calls (F(1,106) = 3.84, n.s.). Also, reputation had no effect on the cat death rate (positive vs. negative reputation: 61 vs. 55%, χ(1) = 0.457, n.s.).

Effect of Capability Demonstration

The above described GLM analysis with the five trust ratings as within-subjects factor and the two experimental factors (reputation and capability demonstration) as between-subjects factors revealed a significant main effect of capability demonstration (F(1,106) = 20.03, p < 0.001, η^2 = 0.159) and also a significant interaction effect between capability demonstration and experience (i.e., time) (F(4,424) = 22.61, p < 0.001, η^2 = 0.176), but no significant three-way interaction between reputation, capability demonstration, and experience (F(4,424) = 1.81, n.s.).

The above described multivariate analysis of variance with the two experimental factors (reputation and demonstration) and the trust ratings for the five points of measurements as dependent variables showed significant differences between the two capability demonstrations for three of the five measures (baseline, week 1, and week 3), whereby a positive demonstration resulted in higher trust ratings than the negative demonstration. However, except of the baseline measures, the differences and effect sizes were quite small (positive vs. negative reputation: baseline: M = 6.5 vs 3.5, F(1,106) = 122.59, p < 0.001, η^2 = 0.536; week 1: M = 7.6 vs. 6.8, F(1,106) = 5.92, p = 0.017, η^2 = 0.053; week 3: M = 8.2 vs. 7.3, F(1,106) = 7.01, p = 0.009 , η^2 = 0.062).

The above described GLM analysis with control calls per week as within-subjects factor and the two experimental factors (reputation and capability demonstration) as between-subjects factors revealed no significant main effect of capability demonstration on the number of control calls (F(1,106) = 0.82, n. s.). Also, capability demonstration had no effect on the cat death (positive vs. negative demonstration: 64 vs. 52%, $\chi^2(1)$ = 1.74, n.s.).

Interpretation of Study Findings Regarding the Development of Overtrust

A central aim of our study was to test the expected development of overtrust by simulating the typical dynamics of experience with technology over time. In line with the assumed general paradigm, repeated positive experience with the pet feeding robot leads to a continuous increase in trust and eventually overtrust on a behavioral and attitudinal level for the majority of participants. Reputation and demonstration had less influence and were primarily relevant for trust measured as a baseline, that is, before the participants could gain any personal experience themselves. Thus, in a simplified scheme, demonstration and reputation form relevant factors for the base level of trust. After this, one's positive or negative experience with the intelligent technology determines the further development of trust. If the experience is repeatedly positive, as in the first trials of our study, this results in trust even beyond the level of calibrated trust. **Figure 6** illustrates this.

In general, given our findings about the primary role of subjective experience (i.e., demonstration and personal experience) compared to cognitive insight (e.g., reputation), interventions that relate to the users' subjective experience may appear more helpful than rational persuasion (e.g., "Warning, do not use the system for other purposes than intended"). In the following sections, we discuss our study findings and other cases of overtrust from a wider perspective and highlight additional psychological mechanisms that might explain user behavior, the development of overtrust. Finally, we suggest potential countermeasures and design approaches toward calibrated trust.

GENERAL DISCUSSION

Instant Rewards and Lack of Falsification

A main reason for the creation of overtrust seems to be the predominance of positive short-term feedback. Initial information such as reputation or demonstration is quickly outweighed by short-term rewards and positive experiences. As long as there is no obvious reason to distrust, people follow the more comfortable way, assuming that the robot is reliable. In our study, this resulted in the participants' decision for the safari instead of the control call. Psychologically, this is quite comprehensible. First, it is well known from consumer choice that people have a natural preference for hedonic, experiential options (here: the safari) over pragmatic options (here: the control call), especially if they can find a reason to justify their choice (e.g., Böhm and Pfister, 1996; Okada, 2005). When translating this to

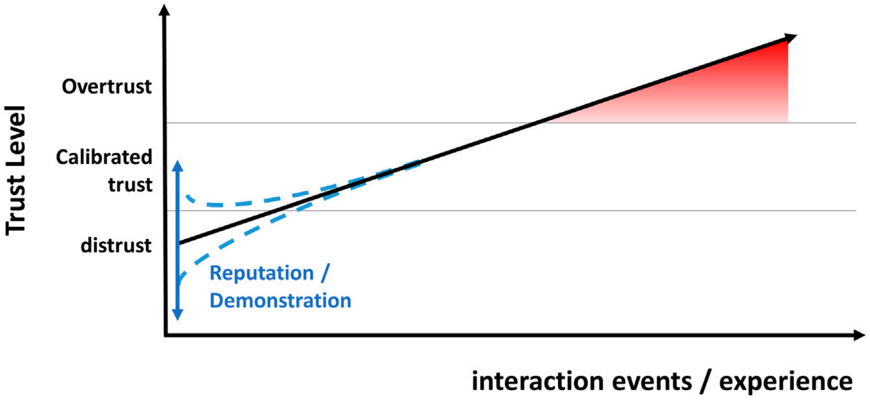


FIGURE 6 | Development of trust beyond system capabilities.

our scenarios, justifications for the hedonic choice may include: the robot performed well so far, why should this change, everybody would have done the same.

Second, people have a general tendency to "test" their assumptions by positivist approaches, searching for confirming information, instead of the more informative contradictory information, the so-called confirmation bias (e.g., Bye, 2012). A classical task to demonstrate this bias is the Wason card task, confronting participants with four cards showing letters or numbers (i.e., A, D, 4, and 7) and a rule about the four cards, namely "if a card has a vowel on one side, then it has an even number on the other side." Participants are then asked which card(s) they need to turn over in order to determine if the rule is true or false. The logically correct answer is to choose A and 7. A is necessary to check whether it has an even number on the other side (otherwise the rule would be falsified); 7 is necessary to check whether it has no vowel on the other side (otherwise the rule would be falsified). However, only about 4% of the participants give this correct answer. The most prominent answer is A and 4, obviously displaying a wish for confirming information. However, in order to retain reliable information about whether an assumption is true or not, one needs to search for situations in which the assumption could possibly be falsified and not situations which are compatible with existing assumption anyway. This is also proposed by Popper's scientific method of falsificationism (e.g., Percival, 2014): Instead of proposing hypotheses and then checking if they can be confirmed by evidence, Popper suggests making conjectures that can potentially be refuted. In the domain of technology this means: if you want to find out about the reliability, you have to confront technology with tasks at the expected limit of its capability.

Inappropriate Generalization and Lack of Differentiation

Another mechanism behind overtrust could be inappropriate generalization from one successful experience to a general capability or, in other scenarios than the pet feeding robot, the lack of differentiation between situations of varying difficulty. As the example of overtrust in the Tesla autopilot showed, people generalize from positive experience in situation A that the system will be able to handle situation B as well. They seem to apply a global concept of trust toward technology similar to that of trust toward humans. Of course, even for humans, a global trust concept does not always hold true (e.g., "My wife is a fantastic driver, I trust she must be a fantastic pilot as well"). But in general, a human might detect what skills from other domains might be transferrable (e.g., "I never played badminton—but it looks a bit like tennis, let's try it with similar moves"), so that trust generalization can to some degree be adequate. For technology, it depends on whether the new situation has been defined beforehand and provides any triggers to activate helpful system skills. Even if a task seems quite "easy" to a human, a robot may not be able to solve it if its algorithms did not define any reaction for it. However, people might lack an exact concept of a technology's capabilities and limitations. If a robot can do stunning things and impress people in one domain, they may see it as a "magician," and readily believe it could do anything. Accordingly, Wagner et al. (2018) already emphasized the importance of mental modeling research and building robots that are more transparent, allowing people to fully understand how the technology will behave.

Transfer of Social Concepts From Human–Human Interaction

As mentioned in the previous section, people tend to transfer concepts from human–human interaction (e.g., the concept of global trust) to human–robot interaction. This tendency also becomes visible in the relative effect of experience vs. reputation. A main finding of our study was that the participants' decision to trust the robot (and the cat death rate) primarily depended on their personal prior positive experience with the robot, whereas the reputation was less relevant. People may follow a rationale of "If I personally have experienced the robot to perform well so many times, it won't let me down the next time." On the contrary, others' shared experiences about the robot's performance were not crucial. Interestingly, this pattern parallels a typical and sensible behavior from human–human interaction: The reliance on personal experience for attitude formation. Even if others tell me about a person, I will build my own opinion based on my own experience. Even though others think that a person is not trustworthy, my relationship to this person can be a different one. I might have a special connection with this person and trust that he or she will never disappoint me. The same counts vice versa: others may have experienced the person as trustworthy, but I have not. Our findings suggest that people may transfer a learnt and sensible behavioral pattern from human–human interaction to human–computer interaction. In parallel to previous studies, showing that people often transfer behavioral patterns from human–human interaction (e.g., rules of courtesy, self-serving attribution biases, and group conformity) to the interaction with computers (e.g., Kiesler and Goetz, 2002; Goetz et al., 2003; Robins et al., 2004; Syrdal et al., 2007), participants behave toward the robot as if the robot had a personal relationship with them and might be more reliable for them than for others. Consequently, they disregard the valuable information they could get from others' experience reports.

Wishful Thinking

Finally, wishful thinking may also play a role for the phenomenon of overtrust. As we know from everyday experience in many contexts, people often do not want to hear about negative aspects or potential risks, given that this would question the current comfortable way of usage. This may pertain to individual behaviors such as the risks of smoking or unhealthy nutrition but also risks on a global level such as nuclear energy, where many people do not want to hear the technology could fail. In fact, the discussion about nuclear energy could be interpreted in parallel to the partly irrational behavior as it appeared in our study: Reputation has no effect at all: In spite of the scientific and media reports about the dangers of nuclear energy, people "trust" it will never fail. Demonstration has a temporary effect: Briefly after the nuclear disasters in Chernobyl and Fukushima, governments around the world decided to ban this technology, but as the memory faded only a few years later, these decisions started to crumble as well. Experiences with the machine dominate other information: In the everyday operation of nuclear power plants worldwide, the positive experience (no direct emissions and plenty of supposedly "clean" energy) by far outweighs the knowledge about the imminent dangers, which creates a widely positive attitude and a flourishing nuclear industry.

Hence, one may question whether it is genuine trust in the technology or to some degree wishful thinking which makes many people still consider nuclear energy a safe technology. In fact, wishful thinking may be particularly pronounced, if people feel that there is no alternative to trusting the technology (e.g., a lack of convincing alternatives to nuclear energy at a large scale). Wishful thinking can also function as a way of dissonance reduction. As dissonance theory (Festinger, 1957) assumes, people strive for conformity between their attitudes, beliefs, and behaviors. If a conflict or dissonance occurs, they typically alter one of the elements. For example, if I do not want to give up smoking, I may alter my belief from "smoking is unhealthy" to "it has never been fully proven that smoking is unhealthy, actually many smokers get quite old" etc. Regarding our study scenario of trusting a pet feeding robot for being able to enjoy a safari trip, a similar mechanism could. If I do not want to change my behavior (e.g., go to the safari instead of doing the boring way to town to make a control call) I better adjust my beliefs (e.g., I trust that the robot is 100% reliable and there is no risk for my pet).

Overtrust From a Phenomenological Perspective and Implications for Design

In the end, this leads to a quite academic discussion whether it is actually trust, altered beliefs, wishful thinking, or any similar factor which is the driving force behind overtrust and risking a technology's failure. On a phenomenological level, all these forces may affect behavior in the same way as genuine trust. This is why we consider it helpful to use the term overtrust in a wider sense for all cases in which people apply a technology beyond the limit of its capability or reliability. If we know that people are prone to the psychological mechanisms discussed above, this implies opportunities but also increased responsibilities for design. The more impressive and overwhelming the technological advancements in various domains, the more difficult it becomes for people to imagine what technology can do or cannot do, and to adequately assess a system's capabilities and limits. Designers must find ways for how a system effectively communicates its features and limits. As discussed under the notion of explainable AI (Monroe, 2018), designers have an ethical responsibility to ensure that their systems explain their strengths and weaknesses to the users and justify their suggestions and decisions in order to prevent unjustified projections and inappropriate trust. In many contexts of HCI design, using psychological mechanisms is actually helpful, for example, using metaphors or designing computer dialogs in parallel to dialogs in human–human interaction. On the other hand, design needs to foresee potential problems resulting from this transfer process and make sure that people do not transfer concepts against their own interest, for example, interpreting an autopilot in parallel to a human driver, which can easily transfer skills from one situation to others.

A central question is how to avoid overtrust and how to support calibrated trust without educating people to generally mistrust technology. Previous suggestions often described "intelligent" system reactions as a possible solution, for example, robots being able to generate information about the person's attentive state (Böhm and Pfister, 1996). However, in order to widen this perspective, we explore how to counteract overtrust by understanding its psychological foundations, including approaches that might not look like smart system behavior at all. A straightforward way to avoid the development of overtrust could be to prevent exclusively positive experience by (harmless) preprogrammed system failure at regular intervals. If, for example, your intelligent fully automatic coffee machine pours too much water into the coffee cup about every third time you press the espresso button, you probably will not trust the machine and leave the room after starting the coffee. Although the coffee tastes excellent, you would feel the machine is not reliable and you better have an eye on it. Of course, this approach of preprogrammed system failure is questionable for several reasons. It causes unneeded difficulties for the user and unneeded negative reputation for the manufacturer. Another, probably more realistic approach could be to work with implicit cues of imperfection (e.g., imperfect grammar in dialog systems), reminding the user that the technology does not work as accurately as the user may assume. From a psychological perspective, such little quirks may even make it appear more human and likable. As revealed in previous research on peoples' relationships with their technical products, a little friction in system interaction is even interpreted as a part of a positive relationship and forgiving statements such as "It [the smartphone] behaves like a modest, loyal servant also be a bit funny—sometimes a program doesn't work properly—it's not the perfect support. But little quirks also make it more likable, more humane" (Chris, cited after Diefenbach and Hassenzahl, 2019, p. 11).

LIMITATIONS AND FUTURE RESEARCH

At least four basic limitations need to be considered for the interpretation of our findings. The first and most general limitation refers to the nonrepresentative sample of participants, that is, rather young people within a limited age range, most of them having an academic background. Although there is no obvious indication that overtrust should be less frequent among older or nonacademic samples, future studies should include more diverse samples of participants.

The second aspect refers to the study's external validity and quantitative focus. Participants' decision to trust the robot or not could realistically affect their emotional experience (i.e., seeing jungle pictures and interesting facts when trusting the robot or an annoying car trip video and feeding results when not trusting the robot) but the risk related to trusting the robot (i.e., the cat dies) was only fictional. Hence, one could question whether the participants would have made the same choice if their real pet's life was in danger. Also, our study was focused on quantitative measures of trust and there was no qualitative assessment of the participants' subjective feelings and how they experienced the scenario. It should be noted, however, that the main aim of our research was the exploration of the assumed paradigm of overtrust and possible additional influencing factors. Even though the general trust rates might have been slightly different if studied in a real-life setting, there is no obvious reason to assume that this would have changed the relative effect of the influencing factors experience, reputation, and demonstration. Future studies should include field studies and complement quantitative accounts with qualitative approaches.

Third, our study was limited to three influencing factors of overtrust (personal experience, reputation, and demonstration) which we identified as dominant in the literature and our pre-study. Hence, while our model provides a valid starting point and framework for the study of overtrust, future research should extend this by an exploration of further influencing factors such as the different social and psychological mechanisms discussed above. Integrating such factors in future research will provide a more holistic picture of the phenomenon of overtrust, its consequences, and potential interventions.

Fourth, one may argue that the course of robot experience our study design provided (i.e., the robot performs well for repeated times and then suddenly fails) was not very realistic, especially given that failures are in themselves rather unlikely. One might even argue that our study design was "unfair" since everybody would trust a machine that has proven so many times. However, even if rarely, the same pattern may occur in real-life scenarios: the cases in which technology fails are rare and experienced only by single users. As a consequence, the most common experience (e.g., watching many people on YouTube doing funny things and taking their hands off the wheel while using the Tesla autopilot and no accident happens) does not reflect the associated risk of blindly trusting the technology. The same happens in other situations without technology being involved, for example, skiing in an avalanche risk area without any problems for many times and then getting killed one day. Above all, these repeated positive experiences make people develop inappropriate trust in a technology or situation and this is what we wanted to simulate. In sum, we created a highly artificial scenario with high internal validity, however, connected to limitations regarding external validity.

In addition to these specific limitations, future work could also further explore the connections to other psychological concepts listed in the discussion section such as inappropriate generalization or cognitive dissonance.

CONCLUSION

As shown by the discussion above, the development of inappropriate trust in intelligent systems has to be seen not as the exception but as the rule. This presents a serious problem when it comes to sensitive domains in which lives or personal well-being might be at stake. The presented case study and psychological analysis make the underlying mechanisms comprehensible, yet they do not deliver any obvious general solutions. The challenge to design and develop technologies in such a way that they prompt an adequate or calibrated level of trust will remain one of the most pressing ones, as long as we are not in a position to develop systems which justify the great amount of trust they are met with by working perfectly.

DATA AVAILABILITY STATEMENT

The data is available under https://doi.org/10.5282/ubm/data.196.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethikkommission der Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universität München (LMU). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AB und DU discussed the pre-study, DU and SD conceived and planned the main study. DU carried out the studies and performed the data analysis. All authors discussed the results and contributed to the final manuscript.

FUNDING

Part of this research has been funded by the German Federal Ministry of Education and Research (BMBF), project GINA (FKZ: 16SV8097).

ACKNOWLEDGMENTS

We thank Verena Brade for the help in conducting the pre-study and online study and Caroline Eckerth for the valuable comments regarding the discussion of socio-psychological concepts.

REFERENCES

Albright, J., Bell, A., Schneider, J., and Nyce, C. (2015). Marketplace of change: automobile insurance in the era of autonomous vehicles. KPMG. Available at: https://assets.kpmg/content/dam/kpmg/pdf/2016/06/id-market-place-of-change-automobile-insurance-in-the-era-of-autonomous-vehicles.pdf (Accessed April 2, 2021).

Autobild (2018). Autonomes Fahren: crash eines Tesla Model X [Online]. Available: https://www.autobild.de/artikel/autonomes-fahren-crash-eines-tesla-model-x-10516973.html. (Accessed August 5, 2019).

Bagheri, N., and Jamieson, G. A. (2004). "The impact of context-related reliability on automation failure detection and scanning behaviour," in Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics, Melbourne,AU, October 17–20, 2004 (IEEE), 212–217.

Barber, B. (1983). *The logic and limits of trust*. New Brunswick, NJ: Rutgers University Press.

Bartneck, C., Kanda, T., Mubin, O., and Al Mahmud, A. (2009). Does the design of a robot influence its animacy and perceived intelligence?. *Int. J. Soc. Robotics.* 1 (2), 195–204. doi:10.1007/s12369-009-0013-7

Bartneck, C., Van Der Hoek, M., Mubin, O., and Al Mahmud, A. (2007). "Daisy, daisy, give me your answer do!" switching off a robot," in Proceedings of the 2nd ACM/IEEE international Conference on human-robot interaction (HRI), Arlington, VA, March 9–March 11, 2007 (IEEE), 217–222.

Böhm, G., and Pfister, H.-R. (1996). Instrumental or emotional evaluations: what determines preferences?. *Acta Psychol (Amst).* 93 (1–3), 135–148. doi:10.1016/0001-6918(96)00017-0

Bye, J. K. (2012). Psychology classics: Wason selection task(Part I) [Online]. Available: https://www.psychologyinaction.org/psychology-in-action-1/2012/10/07/classic-psychology-experiments-wason-selection-task-part-i. (Accessed August 5, 2019).

Carlson, M. S., Drury, J. L., Desai, M., Kwak, H., and Yanco, H. A. (2014). Identifying factors that influence trust in automated cars and medical diagnosis systems. *AAAI spring symposium series*, Palo Alto, CA.

Cathcart, M. E. P. (1997). *The media equation: how people treat computers, television, and new media like real people and places*. New York, NY: Cambridge University Press, 305.

Chen, J. Y., Barnes, M. J., and Harper-Sciarini, M. (2010). *Supervisory control of unmanned vehicles*. Maryland: Aberdeen Proving Ground Army Research Lab.

Clinic, M. (2019). Robotic surgery [Online]. Available: https://www.mayoclinic.org/tests-procedures/robotic-surgery/about/pac-20394974. (Accessed February 3, 2021).

Coeckelbergh, M. (2012). Can we trust robots?. *Ethics Inf. Technol.* 14 (1), 53–60. doi:10.1007/s10676-011-9279-1

Darling, K. (2012). "Extending legal rights to social robots," in Proceedings of the 2012 We robot conference, April 21–22, 2012 (Miami, FL: University of Miami). doi:10.2139/ssrn.2044797

Day, S. N. (2016). Tesla's model S Autopilot is amazing! [Online]. Available: https://www.youtube.com/watch?v=UgNhYGAgmZo. (Accessed August 5, 2019).

De Graaf, M. M. A., and Allouch, S. B. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics Autonomous Syst.* 61 (12), 1476–1486. doi:10.1016/j.robot.2013.07.007

Diefenbach, S., and Hassenzahl, M. (2019). Combining model-based analysis with phenomenological insight: a case study on hedonic product quality. *Qual. Psychol.* 6 (1), 3–26. doi:10.1037/qup0000096

Feil-Seifer, D., and Mataric, M. J. (2005). "Defining socially assistive robotics," in Proceedings of the 9th international conference on rehabilitation robotics, Chicago, June 28–July 1, 2005 (IEEE), 465–468.

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row Peterson.

Folger, J. (2019). Automated trading systems: the pros and cons [Online].Available: https://www.investopedia.com/articles/trading/11/automated-trading-systems.asp. (Accessed May 12, 2019).

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. (2006). *Cognitive Neuroscience. The biology of the mind*. New York, NY: Norton.

Giesen, C. (2016). Auf autopilot [online]. Available: https://www.sueddeutsche.de/wirtschaft/tesla-auf-autopilot-1.3163214. (Accessed September 15, 2016).

Goetz, J., Kiesler, S., and Powers, A. (2003). "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in Proceedings of the 12th IEEE international workshop on robot and human interactive communication, California, October 31–November 2, 2003 (IEEE), 55–60.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Hum. factors.* 53 (5), 517–527. doi:10.1177/0018720811417254

Hoffman, R. R., Lee, J. D., Woods, D. D., Shadbolt, N., Miller, J., and Bradshaw, J. M. (2009). The dynamics of trust in cyberdomains. *IEEE Intell. Syst.* 24 (6), 5–11. doi:10.1109/mis.2009.124

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* 4 (1), 53–71. doi:10.1207/s15327566ijce0401_04

Johns, J. L. J. L. (1996). A concept analysis of trust. *J. Adv. Nurs.* 24 (1), 76–83. doi:10.1046/j.1365-2648.1996.16310.x

Keijsers, M., and Bartneck, C. (2018). "Mindless robots get bullied," in Proceedings of the 2018 ACM/IEEE international Conference on human-robot interaction, New York, NY, March 5–8, 2018, 205–214.

Kiesler, S., and Goetz, J. (2002). "Mental models of robotic assistants," in Proceedings of the CHI'02 extended abstracts on human Factors in computing systems, New York, NY, April 20–25, 2002 576–577.

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum Factors.* 46 (1), 50–80. doi:10.1518/hfes.46.1.50_30392

Lee, K. M., and Nass, C. (2003). "Designing social presence of social actors in human computer interaction," in Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, April 5–10, 2003, 289–296.

Lohani, M., Stokes, C., McCoy, M., Bailey, C. A., and Rivers, S. E. (2016). "Social interaction moderates human-robot trust-reliance relationship and improves stress coping," in Proceedings of the 11th ACM/IEEE international Conference on human-robot interaction (HRI), Christchurch, NZ, March 7–10, 2016 (IEEE), 471–472.

Luhmann, N. (2018). *Trust and power*. Chichester: John Wiley & Sons.

Martelaro, N., Nneji, V. C., Ju, W., and Hinds, P. (2016). "Tell me more: designing HRI to encourage more trust, disclosure, and companionship," in Proceedings of the 11th ACM/IEEE international conference on human-robot interaction, Christchurch, NZ, March 7–10, 2016 (IEEE), 181–188.

Merriam-Webster-Dictionary 2019). Definition of reputation [Online]. Available: https://www.merriam-webster.com/dictionary/reputation. (Accessed May 16, 2021).

Monroe, D. (2018). AI, explain yourself. *Commun. ACM.* 61 (11), 11–13. doi:10.1145/3276742

Moorman, C., Deshpande, R., and Zaltman, G. (1993). Factors affecting trust in market research relationships. *J. mark.* 57 (1), 81–101. doi:10.2307/1252059

Mosier, K. L., Palmer, E. A., and Degani, A. (1992). Electronic checklists: implications for decision making. in *Human factors and ergonomics society annual meeting*. California: Sage Publications, 7–11.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man-Machine Stud.* 27 (5–6), 527–539. doi:10.1016/s0020-7373(87)80013-5

Nass, C., Steuer, J., and Tauber, E. R. (1994). "Computers are social actors," in Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, April 24–28, 1994 ACM, 72–78.

Okada, E. M. (2005). Justification effects on consumer choice of hedonic and utilitarian goods. *J. marketing Res.* 42 (1), 43–53. doi:10.1509/jmkr.42.1.43.56889

Packard, M. G., and Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annu Rev Neurosci.* 25 (1), 563–593. doi:10.1146/annurev.neuro.25.112701.142937

Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum. Factors.* 39 (2), 230–253. doi:10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. *J. Cogn. Eng. Decis. making.* 2 (2), 140–160. doi:10.1518/155534308x284417

Percival, R. S. (2014). Confirmation versus falsificationism. *encyclopedia Clin. Psychol.* doi:10.1002/9781118625392.wbecp388

Qiu, L., and Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: a social relationship perspective to designing information systems. *J. Manag. Inf. Syst.* 25 (4), 145–182. doi:10.2753/mis0742-1222250405

Reek, F. (2015). Autopilot im Tesla Model S: einfach mal loslassen. Available: https://www.sueddeutsche.de/auto/autopilot-im-tesla-model-s-einfach-mal-loslassen-1.2723051. (Accessed November 6, 2016).

Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *J. Personal. Soc.* 49 (1), 95. doi:10.1037/0022-3514.49.1.95

Robinette, P., Howard, A. M., and Wagner, A. R. (2017). Effect of robot performance on human–robot trust in time-critical situationsEffect of Robot Performance on Human-Robot Trust in Time-Critical Situations. *IEEE Trans. Human-Mach. Syst.* 47 (4), 425–436. doi:10.1109/thms.2017.2648849

Robinette, P., Howard, A. M., and Wagner, A. R. (2015). Timing is key for robot trust repair. *International conference on social robotics*. New York, NY: Springer, 574–583.

Robinette, P., Li, W., Allen, R., Howard, A. M., and Wagner, A. R. (2016). "Overtrust of robots in emergency evacuation scenarios", in Proceedings of the eleventh ACM/IEEE international conference on human robot interaction, New Jersey, March 7–10, 2016 (IEEE), 101–108.

Robins, B., Dautenhahn, K., Te Boerkhorst, R., and Billard, A. (2004). "Robots as assistive technology-does appearance matter?", in Proceedings of the 13th IEEE international workshop on robot and human interactive communication, Kurashiki, JP, September 20–22, 2004 (IEEE), 277–282.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *J. PersonalityJ. Pers.* 35 (4), 651–656. doi:10.1111/j.1467-6494.1967.tb01454.x

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Acad. Manag. Review Amr.* 23 (3), 393–404. doi:10.5465/amr.1998.926617

Süddeutsche-Zeitung (2016). Tesla weist Behördenkritik am "Autopilot"-Namen zurück [Online]. Available: https://www.sueddeutsche.de/wirtschaft/auto-tesla-weist-behoerdenkritik-am-autopilot-namen-zurueck-dpa.urn-newsml-dpa-com-20090101-161016-99-828045 [Accessed October 17, 2016].

SAE-International (2018). SAE international releases updated visual chart for its "levels of driving automation,". Available: https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles [Accessed November 12, 2018].

Sanchez, J. (2006). *Factors that affect trust and reliance on an automated aid.* Atlanta, GA: Georgia Institute of Technology.

Sheridan, T. B. (2002). *Humans and automation: system design and research issues.* Santa Monica, CA: Human Factors and Ergonomics Society.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., et al. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PloS one* 1 (1), e39. doi:10.1371/journal.pone.0000039

Sparaco, P. (1995). Airbus seeks to keep pilot, new technology in harmony. *Aviation Week Space Tech.* 142 (5), 62–63.

Syrdal, D. S., Dautenhahn, K., Woods, S. N., Walters, M. L., and Koay, K. L. (2007). "Looking good? Appearance preferences and robot personality inferences at zero acquaintance," in *AAAI Spring symposium: multidisciplinary collaboration for socially assistive robotics*. Palo Alto, CA: AAAI Press, 86–92.

Tesla (2019a). About Tesla [Online]. Available: https://www.tesla.com/about [Accessed May 8, 2019].

Tesla (2019b). Autopilot [Online]. Available: https://www.tesla.com/de_DE/presskit/autopilot [Accessed May 08, 2019].

Ullman, D., and Malle, B. (2016). "The effect of perceived involvement on trust in human-robot interaction," in Proceedings of the the eleventh ACM/IEEE international conference on human robot interaction, New Jersey, March 7–March 10, 2016 (IEEE), 641–642.

Wagner, A. R., Borenstein, J., and Howard, A. (2018). Overtrust in the robotic age. *Commun. ACM.* 61 (9), 22–24. doi:10.1145/3241365

Wagner, A. R. (2009). *The role of trust and relationships in human-robot social interaction.* Atlanta, GA: Georgia Institute of Technology.

Yagoda, R. E., and Gillan, D. J. (2012). You want me to trust a ROBOT? The development of a human–robot interaction trust scaleYou Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale. *Int J Soc Robotics.* 4 (3), 235–248. doi:10.1007/s12369-012-0144-0

Yin, H. H., and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nat Rev Neurosci.* 7 (6), 464–476. doi:10.1038/nrn1919

Zeit-Online (2016). Tesla-Autopilot hielt Lkw für Verkehrsschild [Online]. Available: https://www.zeit.de/mobilitaet/2016-07/autonomes-fahren-tesla-unfall-model-s-autopilot-software [Accessed May 8, 2019].

Zuboff, S. (1989). *The age of the smart machine: the future of work and power.* New York, NY: Basic Books.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 26 May 2021
doi: 10.3389/frobt.2021.599755

Smiles as a Signal of Prosocial Behaviors Toward the Robot in the Therapeutic Setting for Children With Autism Spectrum Disorder

SunKyoung Kim[1], Masakazu Hirokawa[1], Soichiro Matsuda[2], Atsushi Funahashi[3] and Kenji Suzuki[1]*

[1] Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Japan, [2] Faculty of Human Sciences, University of Tsukuba, Tsukuba, Japan, [3] Faculty of Sport Science, Nippon Sport Science University, Yokohama, Japan

OPEN ACCESS

Edited by:
Andreas Riener,
Technische Hochschule Ingolstadt,
Germany

Reviewed by:
Levern Currie,
Virginia Tech, United States
Myounghoon Jeon,
Virginia Tech, United States

****Correspondence:***
SunKyoung Kim
kim@ai.iit.tsukuba.ac.jp

Specialty section:
This article was submitted to
Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

Received: *28 August 2020*
Accepted: *08 March 2021*
Published: *26 May 2021*

Citation:
Kim S, Hirokawa M, Matsuda S, Funahashi A and Suzuki K (2021) Smiles as a Signal of Prosocial Behaviors Toward the Robot in the Therapeutic Setting for Children With Autism Spectrum Disorder. Front. Robot. AI 8:599755. doi: 10.3389/frobt.2021.599755

We explored how robot-assisted therapy based on smile analysis may facilitate the prosocial behaviors of children with autism spectrum disorder. Prosocial behaviors, which are actions for the benefit of others, are required to belong to society and increase the quality of life. As smiling is a candidate for predicting prosocial behaviors in robot-assisted therapy, we measured smiles by annotating behaviors that were recorded with video cameras and by classifying facial muscle activities recorded with a wearable device. While interacting with a robot, the participants experienced two situations where participants' prosocial behaviors are expected, which were supporting the robot to walk and helping the robot from falling. We first explored the overall smiles at specific timings and prosocial behaviors. Then, we explored the smiles triggered by a robot and behavior changes before engaging in prosocial behaviors. The results show that the specific timing of smiles and prosocial behaviors increased in the second session of children with autism spectrum disorder. Additionally, a smile was followed by a series of behaviors before prosocial behavior. With a proposed Bayesian model, smiling, or heading predicted prosocial behaviors with higher accuracy compared to other variables. Particularly, voluntary prosocial behaviors were observed after smiling. The findings of this exploratory study imply that smiles might be a signal of prosocial behaviors. We also suggest a probabilistic model for predicting prosocial behaviors based on smile analysis, which could be applied to personalized robot-assisted therapy by controlling a robot's movements to arouse smiles and increase the probability that a child with autism spectrum disorder will engage in prosocial behaviors.

Keywords: smile, prosocial behavior, robot-assisted therapy, NAO, Bayesian model, electromyogram

1. INTRODUCTION

Robotics have advanced and interactive robots have begun to be made available for a variety of purposes. Accordingly, some researchers have explored the benefits of using robots in therapeutic settings for children with autism spectrum disorder (ASD) (Cabibihan et al., 2013; Huijnen et al., 2018). As the main characteristic of ASD includes a lack of social skills (APA, 2013), robots have been applied in social contexts to facilitate fundamental behaviors for communicating and interacting with others (Pennisi et al., 2016). Researchers have reported that children with ASD show improved behaviors, such as increased eye contact and imitation while interacting with robots (Zheng et al., 2016; Cao et al., 2019). However, the ways that interactions with robots increase specific behaviors of children with ASD has not been fully investigated, and past studies have been limited to targeting basic social skills and behaviors. Therefore, to examine some ways robots may have further therapeutic potential for children with ASD, we designed this novel exploratory study. Smiling was used as a measurable signal of behavior change in therapeutic settings for children with ASD to investigate how robot-assisted therapy may facilitate prosocial behaviors based on smile analysis.

There are possible advantages to including the analysis of smiles in robot-assisted therapy for children with ASD. First, smiling is an innate nonverbal behavior (Shrout and Fiske, 1981; Rashotte, 2002; Parlade et al., 2009). An infant's first involuntarily smiles using mouth corners can be seen during the neonatal period. In the fourth week following birth, they can smile actively by moving muscles around their lips and eyes (Sroufe and Waters, 1976; Messinger et al., 1999). The contractions of specific facial muscles—the orbicularis oculi and zygomaticus major—have been observed when infants, as well as adults, are in a good mood. This muscle activity is accompanied by changes around the lips and eyes (Frank et al., 1993; Parlade et al., 2009). Although children with ASD have difficulty recognizing the smiles of others, they can exhibit voluntary smiles using those muscles (Hermelin and O'Connor, 1985; Sato, 2017).

Moreover, smiles can provide social and emotional information (Rashotte, 2002; Martin et al., 2017). The meanings of smiles differ depending on social situations, and the interpretation of other behaviors before, during, or after smiles can vary (Messinger et al., 2001). For instance, smiling when talking about positive things can be explained differently than smiling when talking about negative things (Sonnby-Borgström, 2002). This characteristic of smiles provides additional information for understanding other behaviors. Also, smiles may provide a criterion for evaluating the current developmental stage and progress in children with ASD (Funahashi et al., 2014; Samad et al., 2018).

Lastly, smiles may be a predictor of positive behaviors. Prosocial behaviors are actions that can benefit others, such as helping (Warneken and Tomasello, 2009), cooperating (Brownell, 2013), sharing resources (Dunfield, 2014), or providing emotional support (Svetlova et al., 2010). In previous studies, prosocial behaviors have been investigated in combination with positive moods (Carlson et al., 1988; Guéguen and De Gail, 2003; Telle and Pfister, 2016), and smiles were considered as an indicator of positive mood (Cunnigham, 1979; Baron, 1997; Forgas, 2002; Drouvelis and Grosskopf, 2016). Participants in the studies were willing to pick up a dropped pen, give change for a dollar, and play a game cooperatively after smiling. These findings suggest that people tend to engage in prosocial behaviors after they smile.

Learning prosocial behaviors is important for all children. Considering the personal advantages of receiving help from others and the social benefits of engaging in prosocial behaviors toward others, it is necessary for children with ASD to develop prosocial behaviors. Although the developmental sequence and timing have varied in previous studies, it has been reported that children with ASD can demonstrate prosocial behaviors. Action-based prosocial behaviors, such as picking up and returning items someone has dropped, have been observed in children with ASD between 24 and 60 months of age (Liebal et al., 2008). Also, emotion-based prosocial behaviors, such as responding to others' negative emotions, were reported in a study of 6- and 7-year-old children with ASD (Deschamps et al., 2014). As each child with ASD is in a different social developmental stage (APA, 2013), children with ASD need to practice various prosocial behaviors individually.

Robots could provide personalized therapy for children with ASD. Improvised interactions using a teleoperation method were applied to robot-assisted therapy (Thill et al., 2012; Hirokawa et al., 2018). In those studies, a robot's movements were controlled depending on a child's responses. In this research, we teleoperated a small humanoid robot called NAO (SoftBank Robotics Corp., Paris, France) and observed child–robot interactions in a therapeutic setting. The NAO robot is known for its use in education and therapy (Diehl et al., 2012; Ismail et al., 2012; Bharatharaj et al., 2017; Huijnen et al., 2017). In particular, it can play various roles as either a trainer or a peer of children with ASD. Additionally, using a NAO in the role of care-receiver in the classroom has been suggested for 3- to 6-year-old children to help them learn new words (Tanaka and Matsuzoe, 2012). Therefore, we assigned the care-receiving role to a NAO robot to create a social context in which children with ASD can practice prosocial behaviors. We sought to examine the children's behaviors with the robot when it walked around or fell down. We assumed that each child might smile before engaging in prosocial behaviors.

Thus, the purpose of this study was to explore the potential of personalized robot-assisted therapy based on smile detection for facilitating prosocial behaviors in therapy. The research was guided by the following research question:

Q. Are smiles a potential key factor in predicting prosocial behaviors in walking and falling situations with a robot?

To explore the research question, we adopted video analysis and a physiological signal-based method in a therapeutic setting; participants included children with ASD and typically developing (TD) children. The data obtained regarding TD children were used when observing the behavioral patterns of children with ASD to determine if they are the same. We first measured the duration of smiles and prosocial behaviors through video observation. Second, we complemented data regarding unobserved smiles with electromyogram (EMG) data from each participant. Third, we observed changes in smiles and prosocial behaviors. Finally, we applied a Bayesian framework with conditional probability to explore the potential of smiling as a predictive factor of prosocial behaviors as follows: If the occurrence of prosocial behaviors changes when smiles appear and the Bayesian model shows that smiles have predictability potential, this exploratory study may suggest a new framework for personalized robot-assisted therapy.

2. MATERIALS AND METHODS

2.1. Participants

For this exploratory study, we recruited six children identified as having mild to moderate levels of ASD through the Institute for Developmental Research at the Aichi Human Service Center in Japan. For comparison, six TD children were also recruited. Children with ASD participated in four sessions, and TD children participated in three sessions of robot-assisted activities directed by a therapist. However, we were not able to include all the sessions due to the limitations involved with making the robot fall. Hence, we employed two sessions of children with ASD and one session of TD children; all of the sessions included both the robot walking situation and the robot falling situation. The average age of six children with ASD (four boys and two girls) was 9.67 years old (6–16, *SD* = 3.50) and the average age of six TD children (three boys and three girls) was 9.83 years old (6–11, *SD* = 2.04). None of the 12 children indicated they had any concerns about interacting with a robot and wearing a device. **Table 1** shows the age information of the participants included in each session.

This research was approved by the Ethical Committee based on the Declaration of Helsinki and ethical rules established by the Aichi Human Service Center. The research data were collected in an intervention room of the same institute in compliance with the ethical principles. All caregivers of the children agreed to written informed consent and participated in the entire session.

2.2. Robot

A NAO robot was adopted to create social situations. It is a small-sized (58 cm in height) humanoid robot. NAO has been applied for therapy, rehabilitation, and education contexts requiring interactions with humans (Ismail et al., 2012; Tanaka and Matsuzoe, 2012; Pulido et al., 2017). It can communicate by expressing verbal and nonverbal behaviors. The 26 joints in the head, arms, legs, and pelvis of an NAO robot enable it to perform various motions, such as walking, sitting, and grasping. However, the movements are inflexible and unbalanced compared to human peers, which could lead children to perceive the robot as a care-receiver. After considering the functions and limitations of the NAO robot, we chose "walking with the robot" as the social context for this study. The expected social situations in the given scenario were (1) the robot walking, and (2) the robot falling; the desirable prosocial behaviors we looked for from the children were (1) helping the robot to walk, and (2) helping the robot stand up after it fell down. The NAO robot was controlled using teleoperated methods to create real-time interactions. In this study, we used the Wizard of OZ technique, a research method to make participants feel that they are interacting with an autonomous system (Riek, 2012). A human operator observed each child's responses to the NAO robot in the observation room and controlled the robot's movement by following the cues from a therapist in real time. The voice function of the robot was not used to make simplified interactions and to focus on nonverbal behaviors, which can affect prosocial behaviors.

TABLE 1 | Information on participants.

Participant ID	Age	Analyzed session
ASD-P1	16	Session 1, 2
ASD-P2	11	Session 1, 2
ASD-P3	8	Session 2, 4
ASD-P4	8	Session 2, 4
ASD-P5	6	Session 3, 4
ASD-P6	9	Session 1, 2
TD-P1	11	Session 2
TD-P2	11	Session 1
TD-P3	11	Session 3
TD-P4	9	Session 2
TD-P5	6	Session 1
TD-P6	11	Session 1

2.3. Apparatus

To analyze each participant's smiles and behaviors, video cameras and a wearable device, called the Smile Reader, were used in this research (**Figures 1**, **2**). Four video cameras were installed on the ceiling of the intervention room. A therapist traced and captured each participant's movements with a hand-held video camera. The Smile Reader was used to record surface EMG from the facial muscles (Gruebler and Suzuki, 2014). The device was attached to both sides of the participant's face.

We used the wearable device with EMG sensors because it was designed and developed specifically for smile detection (Gruebler and Suzuki, 2014). This device can detect the contractions of facial muscles related to smiling—the orbicularis oculi and zygomaticus major. These facial muscle areas have been researched with EMG sensors to measure specific smiles that

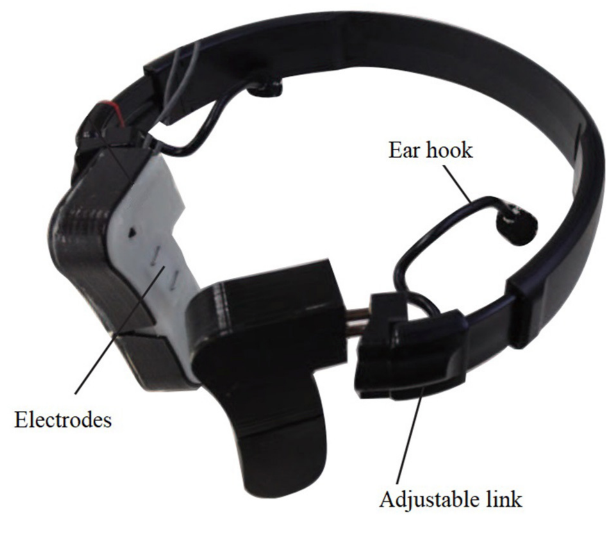


FIGURE 1 | Smile Reader which was used in the robot-assisted therapy.

show spontaneous and positive emotions (Frank et al., 1993; Mauss and Robinson, 2009; Johnson et al., 2010; Perusquía-Hernández et al., 2019). Compared to other physiological sensors, such as electroencephalography and functional MRI, facial EMG can be attached directly to the facial muscles involved in smiling (Maria et al., 2019). Also, it can be used in both laboratory and therapy settings (Hirokawa et al., 2016). Furthermore, the performance evaluation of the Smile Reader has been investigated with adults in a laboratory and children with ASD in therapy; the device has proven reliability for accuracy in smile detection (Funahashi et al., 2014; Gruebler and Suzuki, 2014; Hirokawa et al., 2018).

In this research, each participant's facial EMG was recorded with the Smile Reader including four pairs of active electrodes and a BioLog (S&ME, Japan), a portable EMG logger that includes an amplifier. The devices were connected to a laptop wirelessly, and EMG signals were recorded in real time. To synchronize video and EMG data, a noticeable sign was included in the recorded EMG by using a time tagger.

2.4. Procedure

This exploratory study is based on data collected during robot-assisted therapy. A NAO robot was used to assist a therapist in facilitating prosocial behavior of each child with ASD. The children with ASD participated in this research during the therapy. TD children who joined this research experienced the same procedure. Each child participated in a session every 2–3 weeks, a total of about 3 months. Each session lasted for 20–30 min, and every child was allowed to interact with the robot, a parent, or a therapist without restriction during all sessions. The 9.6 m^2 area where each child could interact with the robot was fenced for safety (**Figure 3**). Their behaviors were recorded by ceiling cameras and the therapist's camera. Each therapy session was divided into four stages, and each stage included a specific cue from the therapist and the corresponding robot behaviors (**Figure 4**). When there were no cues from therapists, a human operator improvised the robot's movements.

The prescribed procedures of each stage are described in sections 2.4.1–2.4.5. The designed or anticipated behaviors and interactions related to the study were examined; improvised behaviors or interactions were excluded from video analysis.

2.4.1. Preparation

Variables of this study were defined as follows: As in prior studies, smiles were defined as changes around the lips or eyes because the facial muscles related to positive affect are contracted by the changes (Frank et al., 1993; Parlade et al., 2009). Prosocial behaviors were defined differently in the two situations. In the walking situation, the children's prosocial behaviors included (1) approaching NAO to hold hand(s) and holding NAO's hand(s) or (2) walking together while holding NAO's hand(s). Prosocial behaviors during the falling of the NAO robot were defined as approaching NAO to hold its body and help the robot stand up.

Before interacting with a NAO robot, each participant was introduced to a preparation room and informed about the wearable device. While wearing the device that records facial muscle activities, each child was asked to watch 20 images that appeared on a computer screen. Each image appeared on the screen for 2 s. The images were emotionally neutral stimuli selected by a medical examiner, and they were used as a baseline to train the artificial neural network (ANN).

2.4.2. Stage 1

Each child moved into an intervention room with a therapist and parent. The first stage began when the therapist pressed a button for a time tagger connected to EMG logging and opened the door of the intervention room. The therapist introduced each child to the robot, and the robot greeted them by moving its arms and turning its head to look around.

2.4.3. Stage 2

In the second stage, each child interacted freely with the robot. In the middle of this stage, the therapist suggested the child and robot play a game of rock-paper-scissors or that they play catch by throwing and catching small beanbags. The NAO robot used hand gestures and body movements for each game. For example, during the rock-paper-scissors game, the robot made a handshape of rock, paper, or scissors, and when the robot won, it raised its arms. When the robot lost the game, it looked down and shook its head from side to side. When playing with the beanbags,

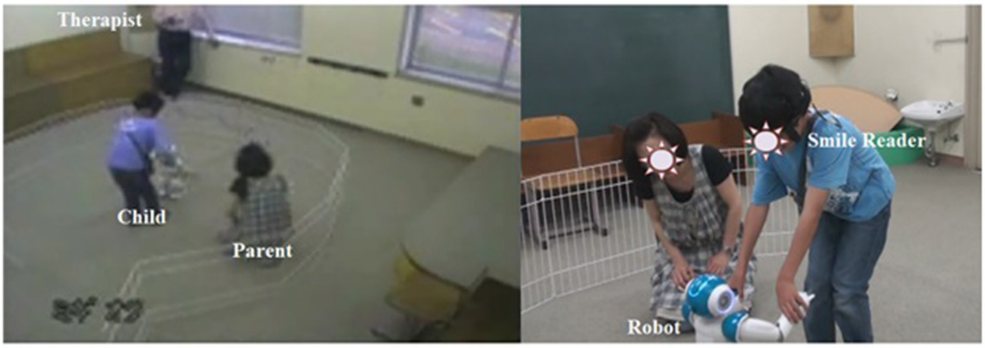


FIGURE 2 | A child wearing a Smile Reader in the intervention room (captured by video cameras).

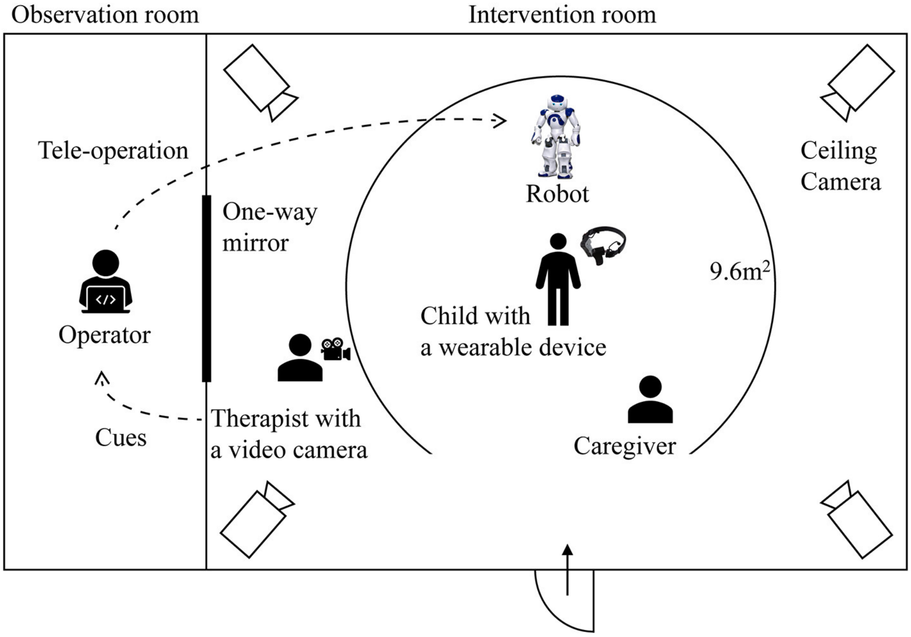
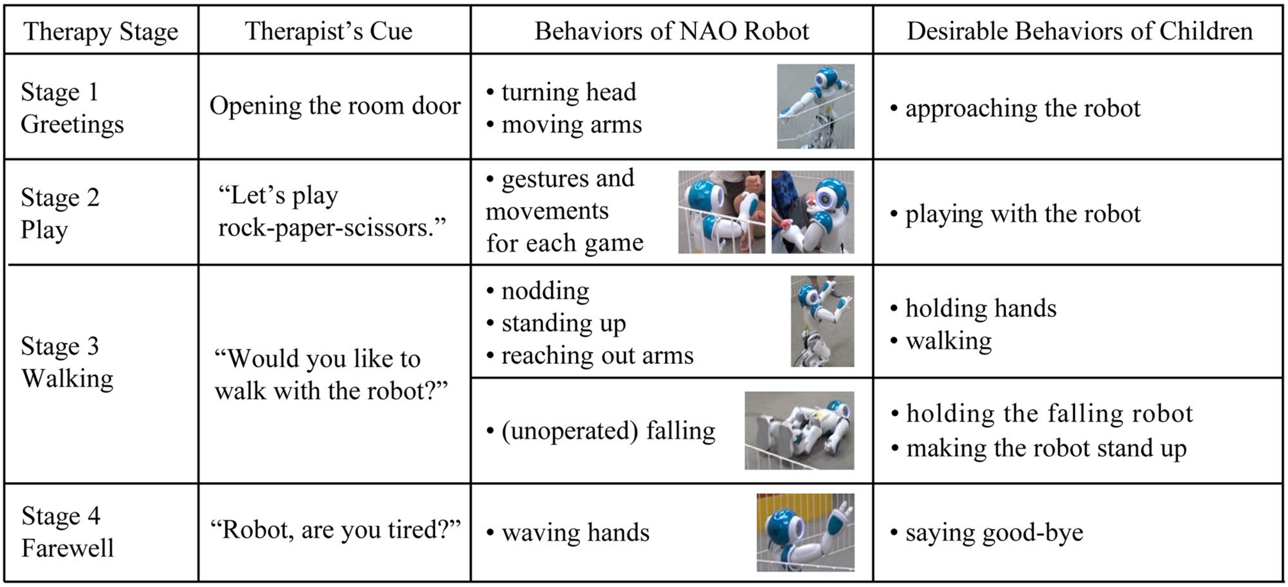


FIGURE 3 | Overview of the experimental setting.



Therapy Stage	Therapist's Cue	Behaviors of NAO Robot	Desirable Behaviors of Children
Stage 1 Greetings	Opening the room door	• turning head • moving arms	• approaching the robot
Stage 2 Play	"Let's play rock-paper-scissors."	• gestures and movements for each game	• playing with the robot
Stage 3 Walking	"Would you like to walk with the robot?"	• nodding • standing up • reaching out arms	• holding hands • walking
		• (unoperated) falling	• holding the falling robot • making the robot stand up
Stage 4 Farewell	"Robot, are you tired?"	• waving hands	• saying good-bye

FIGURE 4 | The designed behaviors of a NAO robot and desirable behaviors of children in each therapy stage.

the robot reached out its hands to receive the beanbags from a child and used its arms to throw them toward the child. Upon failing to catch a beanbag, the robot looked down, raised an arm, and tapped its own head.

2.4.4. Stage 3

In the third stage, the therapist suggested walking together with the NAO robot, and the robot agreed by nodding, standing up, or reaching out with its arms. In this scenario, the desirable behaviors of children included holding the hands of the robot, and walking together. When a child did not show any expected behaviors, the therapist or a parent verbally directed the child to help the robot walk. However, when the robot fell by chance, the therapist observed each child's spontaneous responses without providing direction. The desirable expected behaviors of children were those that helped the robot stand up. When a child helped the robot to walk or stand up, the therapist said, "Thank you" to the child on behalf of the robot.

2.4.5. Stage 4

In the last stage, the therapist suggested finishing the session. In response to the therapist's cue, the NAO nodded and waved a hand. After finishing the last stages, each child moved to the preparation room with a parent and took off the wearable device.

2.5. Video Analysis

Video analysis was adopted to measure the duration of smiles and prosocial behaviors.

2.5.1. Step 1: Annotating Video Streams

To measure smiles and prosocial behaviors, each video was annotated based on the duration of each child's behaviors. The annotation included the beginning of the session, smiles, prosocial behaviors, other remarkable behaviors—such as waving hands, talking, and gesturing—and the unobservable facial expressions of each participant. The duration of smiles and prosocial behaviors were measured per millisecond (ms) by two trained examiners using Dartfish, a tagging software (Dartfish, Fribourg, Switzerland).

In the walking situation, the prosocial behaviors of children included (1) approaching NAO to hold hand(s) and holding NAO's hand(s) or (2) walking together while holding NAO's hand(s). We identified the point when a child started approaching NAO to hold hand(s) as the starting time of the prosocial behavior. Prosocial behaviors during the falling of the NAO robot were defined as approaching NAO to hold the body and helping the robot stand up. When the robot was falling in front of a child, holding the falling robot or making the robot stand up were defined as prosocial behaviors. We identified the point when a child released his or her hold on NAO's hand(s) or body as the ending time of prosocial behavior. The duration of prosocial behavior was calculated as the amount of time between the starting point and the ending point of the behavior. The duration of smiles was calculated as the amount of time between the starting time and ending time of the facial expression with upward lip corners or downward eye corners.

2.5.2. Step 2: Selecting Segments of Video for Analysis

To explore the specific timing of smiles and prosocial behaviors, six segments of the video were selected: (a) 1 min after entering the intervention room (encounter with the robot), (b) 1 min before starting prosocial behaviors in the walking situation, (c) 1 min after starting prosocial behaviors in the walking situation, (d) 1 min before starting prosocial behaviors in the falling situation, (e) the duration of the first smile when the robot is falling down, and (f) 1 min after the robot is adjusted in the falling situation (**Figure 5**). The segments were selected considering specific timings that might affect smiles and prosocial behaviors. Segments (b), (d), and (e) were selected considering that, in previous studies, prosocial behaviors occurred more frequently after smiling (Guéguen and De Gail, 2003; Vrugt and Vet, 2009). Segment (a) was selected considering that first impressions might change how a child behaved toward the robot throughout the session (Willis and Todorov, 2006). To explore if smiles before prosocial behaviors are more related to prosocial behaviors, (c) and (f)—smiles during or after prosocial behaviors—were selected. Each segment length was determined considering the duration of one type of activity, such as greeting the robot or playing rock-paper-scissors. Each activity lasted ∼1 min. Also, the length was determined considering the duration of affect, including both emotion and mood (Beedie et al., 2005; Mauss and Robinson, 2009). As emotion is defined in seconds, shorter than mood, we considered a minimum duration of mood and a maximum duration of emotion.

Moreover, the analyzed timings were limited to the robot's first experience of walking and falling in a session, as each participant experienced a different number and duration of the social situations depending on participated sessions and interactions with the robot.

For the annotation of smiles in selected segments, reliability between the two examiners was high. The average intraclass correlation coefficient was 0.849 with 95% confidence interval from 0.811 to 0.879 [$F_{(307,\ 307)} = 6.629, p < 0.001$].

2.5.3. Step 3: Observing Behavior Changes Before Engaging in Prosocial Behaviors

To explore how behavior changes happen after the robot's movement, 10 s of videos were selected before engaging in prosocial behaviors. The duration was selected by a study related to measuring affective engagement (Rudovic et al., 2017). There were four questions applied to the observations. First, are smiles observed before engaging in prosocial behaviors? Second, if a smile is observed, what triggered smiles? Third, what are the subsequent behaviors with smiles? Fourth, how are those behaviors linked to prosocial behaviors? To explore the questions, the head direction, facial expression, and body movement of each child were annotated every 1 s. The purpose of this observation was to investigate whether smiles can be triggered by a robot's movement and explore whether smiling is a potential predictive factor of prosocial behaviors.

2.5.4. Step 4: Synchronizing EMG and Video Data

All the annotated video parts were synchronized with EMG data. As Stage 1 started by opening the intervention room door, and a therapist logged the moment into EMG data using a time tagger, we first checked the tag. Next, we synchronized the start timing of each annotation and the position of EMG. Lastly, we checked the synchronization in video streams.

2.6. EMG Signal Processing for Estimation of Unobserved Smiles

The smiles presented in this research are complemented by the durations of smiles detected by the EMG signal processing, as there were unobservable smiles. The ratio of unobservable parts in a whole session was a minimum of 2% and a maximum of 25% for a child with ASD, and a minimum of 3% and a maximum of 14% for a TD child. We used the EMG recordings from the wearable device to estimate smiles during the fragments unobservable with the video data. Based on this estimation,

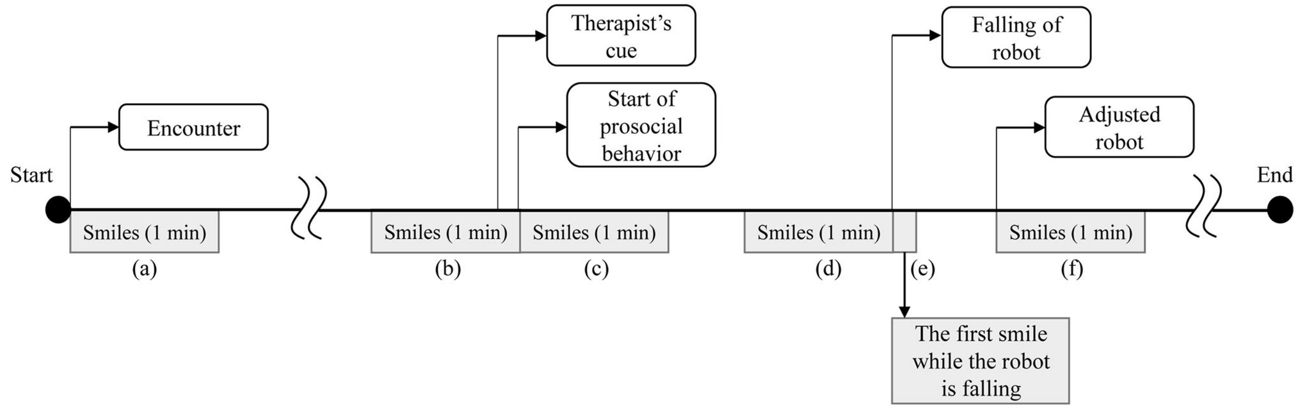


FIGURE 5 | Timeline of analyzed parts of smiles.

the duration of smiles was calculated. All results presented in this research were obtained from combined durations with the observable segments of video data and the unobservable segments with EMG data. We verified that none of the presented trends changed with the estimation of the EMG data.

To estimate unobserved smiles, cross-validation in machine learning was applied. We measured each child's facial EMG signals using four pairs of electrodes. When video cameras could not capture their face because children unexpectedly turned around or stood up, which were frequently included when doing prosocial behaviors, we detected smiles by the following signal processing algorithm. First, a 50–350 Hz band-pass filter was applied to extract the EMG signals by removing noise and outliers. Since each EMG signal is a superposition of multiple facial muscle activities, Independent Component Analysis (ICA) was applied to convert the filtered data into four independent signals to increase the saliency of each signal. Then, root-mean-squared averaging was applied to each independent component with a 100 ms averaging window. Finally, an ANN was trained using the analysis of human coders as a teaching signal to recognize the unobserved smiles of each participant. Among data of smile and no-smile, datasets having less noise and artifacts were used for training to evaluate the predictive performance on the testing set. This signal processing was performed by MATLAB R2017b (Mathworks, USA).

In previous studies, ANN has been used and suggested as a classifier to improve the classification accuracy for EMG signals (Maria et al., 2019; Singh et al., 2020). The performance was different depending on experimental settings. Other classification methods, such as Support Vector Machine and Convolutional Neural Network, were also suggested to increase classification accuracy (Toledo-Pérez et al., 2019; Bakırcğlu and Özkurt, 2020). However, the Smile Reader showed high accuracy with ANN (Gruebler and Suzuki, 2014; Hirokawa et al., 2018). When an ANN was applied to detect positive facial expressions with the Smile Reader, the average Kappa Coefficient between human coders and the classifier was 0.95 (Gruebler and Suzuki, 2014), which shows highly identical inter-rater agreement. Therefore, we applied the ANN classification to detect the unobserved smiles of each participant.

3. RESULTS

The results are organized in three subsections presented below. The first part, *Observation of Different Behaviors*, presents different aspects of participants in specific timings.

The second part, *Observation of Common Behavior Changes*, presents the common behavior changes witnessed before the children engaged in prosocial behaviors. We explored how their behavior changed following the robot's movements.

The third part, the *Behavior Model Framework*, presents the proposed Bayesian model framework for probabilistic inference with the observed variables. We implemented a model based on the data derived from the robot-assisted therapy. This model, however, is not conclusive due to the small sample size of this study, but it is representative of the method we propose that can be applied to similar robot-assisted therapies.

3.1. Observation of Different Behaviors

We observed smiles at specific timings to explore which timings could be more related to prosocial behaviors. Also, we explored whether different behaviors are observed between children with ASD and TD children, and the two sessions of children with ASD.

3.1.1. Smiles and Prosocial Behaviors in the Walking Situation

On average, children with ASD smiled longer than TD children in the walking situation (**Tables 2**, **3**). TD children smiled the most when they entered the intervention room, then smiled less. On the other hand, TD children engaged in prosocial behaviors longer than children with ASD (**Figure 6**).

When comparing the first and second session of children with ASD, each child with ASD showed different changes in the second session. **Figure 7** indicates relationships between the duration of smiles and the duration of each participant's prosocial behaviors in the walking situation. The duration of smiles is the sum of smiles during the encounter and before walking together with the robot, as shown in segments in **Figure 5**a,b, which increased in the second session. The duration of prosocial behaviors is calculated as the sum of helping the robot walk. Empty symbols signify the first session; filled symbols signify the second session. The numbers in the symbols indicate the participant number of each child with ASD. Four children (ASD-P1, ASD-P4, ASD-P5, and ASD-P6) out of six children with ASD showed a longer duration of smiles and longer prosocial behaviors during the second session than during the first session. One child (ASD-P2) showed a shorter duration of smiles and a shorter prosocial behavior during the second session than in the first session. Another child (ASD-P3) showed an increased duration of prosocial behaviors but showed a decreased duration of smiles in the second session. Instead, the child started to sing a song before doing prosocial behaviors. The results imply the possibility of a positive relationship between smiles and prosocial behaviors in children with ASD.

TABLE 2 | The averaged smiles in the first and second session of children with autism spectrum disorder (ASD) (unit is seconds).

ASD session 1 Smile duration Mean ± SD	ASD session 2 Smile duration Mean ± SD	Timing
23.7 ± 17.9	24.9 ± 20.1	1 min after entering the intervention room
20.2 ± 12.9	24.6 ± 11.1	1 min before walking with the robot
35.5 ± 16.8	19.6 ± 15.2	1 min after starting walking with the robot
28.1 ± 11.9	24 ± 10.1	1 min before falling of the robot
7.7 ± 4.3	9.7 ± 4.3	While the robot was falling
43 ± 10.1	17.8 ± 8.7	1 min after the fallen robot was adjusted

TABLE 3 | The averaged smiles in the first session of children with autism spectrum disorder (ASD) and typically developing (TD) children (unit is seconds).

ASD session 1 Smile duration Mean ± SD	TD session 1 Smile duration Mean ± SD	Timing
23.7 ± 17.9	20.4 ± 15.6	1 min after entering the intervention room
20.2 ± 12.9	8.5 ± 4.9	1 min before walking with the robot
35.5 ± 16.8	16.6 ± 16.3	1 min after starting walking with the robot
28.1 ± 11.9	17.7 ± 15.3	1 min before falling of the robot
7.7 ± 4.3	2 ± 3.5	While the robot was falling
43 ± 10.1	10.4 ± 7.8	1 min after the fallen robot was adjusted

3.1.2. Smiles and Prosocial Behaviors in the Falling Situation

On average, children with ASD smiled longer than TD children in the falling situation (**Tables 2**, **3**).

All children with ASD smiled at the robot during the falling moment in the first and second session. Among them, two children with ASD (ASD-P2 and ASD-P6) showed prosocial behaviors in the first session. Three children with ASD (ASD-P1, ASD-P2, and ASD-P5) showed prosocial behaviors in the second session.

In contrast, three TD children (TD-P1, TD-P2, and TD-P5) did not smile while the robot was falling. Among TD children, one child (TD-P4) immediately helped the robot stand up. Two TD children (TD-P2 and TD-P6) helped the robot after watching the fallen robot for ∼10 s.

3.2. Observation of Common Behavior Changes

To investigate how behaviors change after the robot's movements and find a common series of behaviors before engaging in prosocial behaviors, we observed the behaviors of each participant 10 s before prosocial behaviors. If smiles are observable and other behaviors follow the smile, we might be able to predict behaviors after smiles. Also, if smiles are triggered by

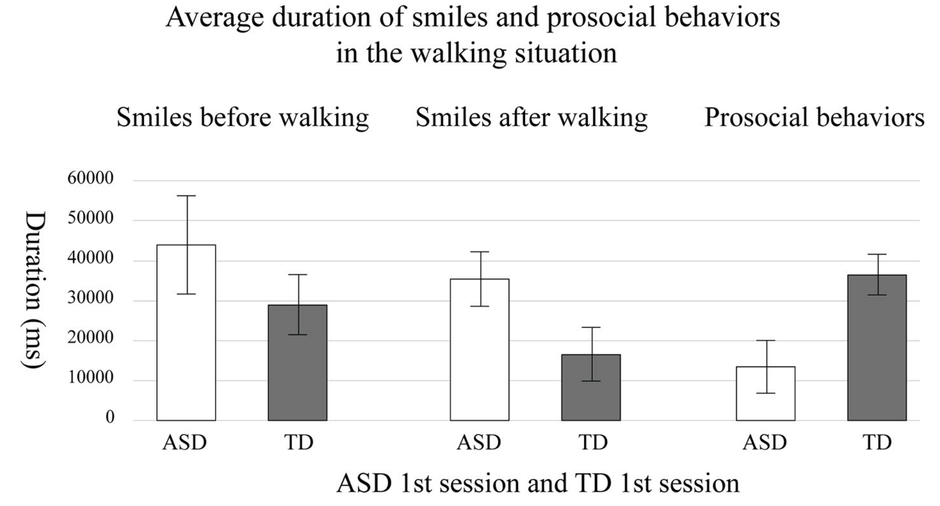


FIGURE 6 | The average duration of the smiles and prosocial behaviors in the first session of children with autism spectrum disorder (ASD) and typically developing (TD) children. The error bar means standard error.

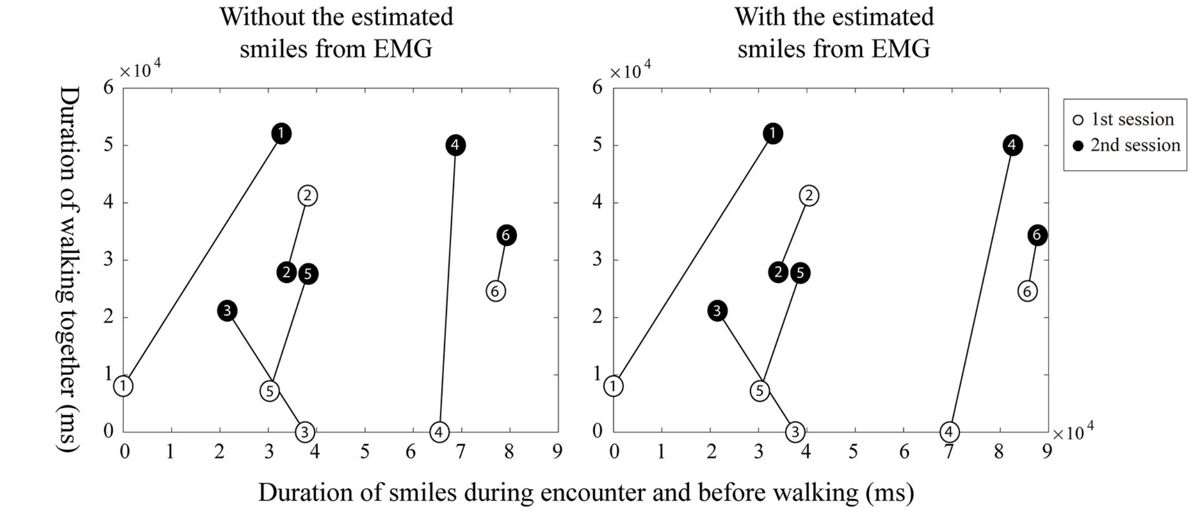


FIGURE 7 | The possible relationships between smiles and prosocial behaviors from each child with autism spectrum disorder (ASD) in the walking situation.

a robot, we might be able to arouse timely smiles and facilitate prosocial behaviors using a robot.

The observation was based on a total of 36 cases of robot walking and falling situations, including 12 cases of children with ASD in the first session, 12 cases of children with ASD in the second session, and 12 cases of TD children in the first session. We observed four types of common cases.

3.2.1. Before Walking of Robot

Case A: Cases of children who showed smiles and prosocial behaviors.

ASD-P1, ASD-P2, ASD-P4, and ASD-P6 showed smiles toward the robot after watching the robot's movements, such as nodding and reaching out its arms. After smiling, they maintained their head direction toward the robot, went closer to the robot, then showed prosocial behaviors voluntarily. In the case of ASD-P6, the child showed the same pattern of behaviors both in the first and second session.

We found similar interactions from TD-P2, TD-P4, and TD-P5. Children, who smiled and maintained their head direction toward the robot, went closer to the robot, and showed prosocial behaviors voluntarily. The smiles were triggered by the robot's movement or observation of interactions between a parent and the robot.

On the other hand, ASD-P2 showed smiles toward the robot after watching the robot's nodding. However, the child's head direction changed to toward their own body, and the child started to move their own fingers without smiling. When the child was focusing on his fingers, his parents tapped his back two times and suggested walking with the robot. The child looked at his parents and then stood up to hold the robot's hands.

Case B: Cases of children who did not show smiles and prosocial behaviors.

ASD-P3 in the first session did not smile after watching the robot's nodding and standing up, and did not show prosocial behaviors. The robot's movements made the child move the head toward the robot temporarily; however, the child did not maintain the head direction. The child looked at the therapist's camera and made a V shape with fingers in the first session.

Case C: Cases of children who showed smiles but did not show prosocial behaviors.

ASD-P4 in the first session smiled toward the robot after watching the robot's standing up. However, the child did not maintain the head direction. The child started to smile toward the parents and went closer to them.

Case D: Cases of children who did not show smiles but showed prosocial behaviors.

Total eight cases from children with ASD and TD children did not smile after watching the robot's movements but showed prosocial behaviors. Before engaging in prosocial behaviors, they received a parent's help or additional direction from the therapist. When their head direction was toward the robot, the child started to follow that direction.

3.2.2. During Falling of Robot

Case A: Cases of children who showed smiles and prosocial behaviors.

In the five cases, children with ASD smiled toward the robot when it was falling and then they moved closer to the robot. Their head direction was continuously directed toward the robot. The children smiled toward the robot before starting to engage in prosocial behaviors.

TD-P2, TD-P4, and TD-P6 also showed smiles and prosocial behaviors. However, they showed different aspects of behaviors that were not observed in children with ASD. TD-P4 and TD-P6 looked at the therapist after doing prosocial behaviors. TD-P2 did not show smiles when the robot was falling. However, the child looked at the therapist after the robot fell and asked the therapist if helping the robot is allowed. Then the child smiled toward the robot before engaging in prosocial behaviors.

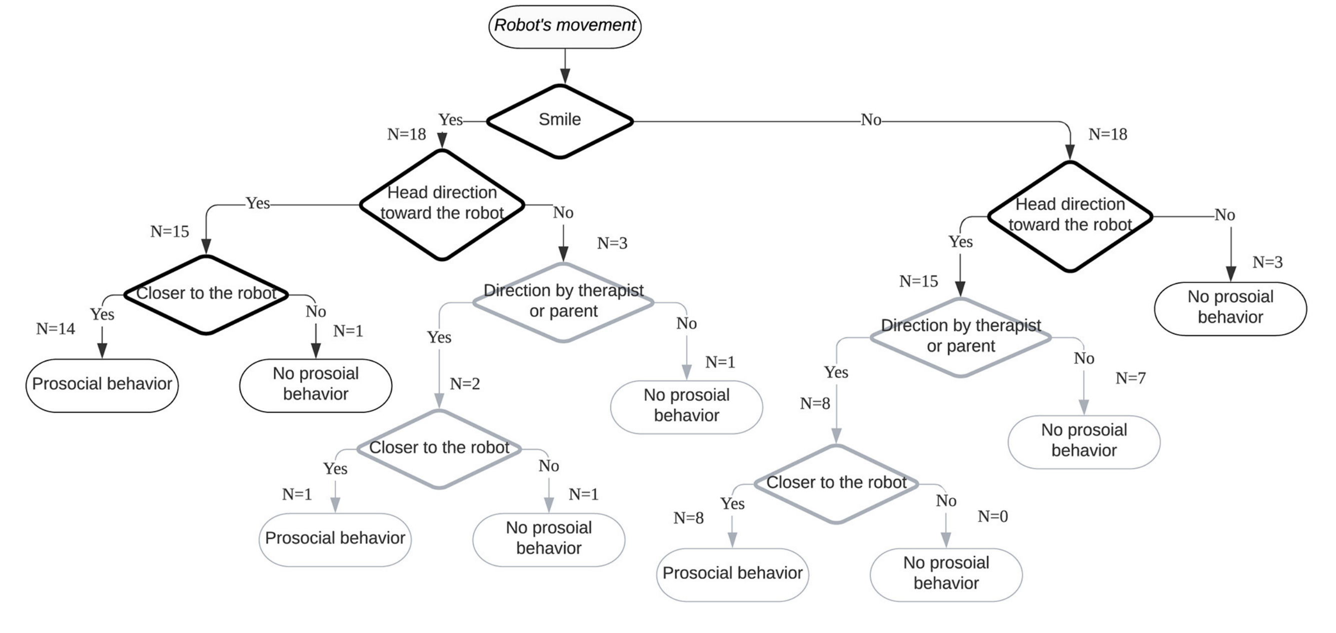


FIGURE 8 | A flowchart of child–robot interactions before prosocial behaviors.

Case B: Cases of children who did not show smiles and prosocial behaviors.

In the six cases, children with ASD released the robot's hands and became distant from the robot when the robot was falling. The head direction was continuously directed toward the robot.

On the other hand, TD-P1 and TD-P3 looked at the therapist after distancing from the robot. TD-P5 watched the robot's falling while sitting behind and holding onto a parent. The head direction of this child was continuously toward the robot, but this child did not show any different facial expressions or body movements after seeing the robot falling.

Case C: Cases of children who showed smiles but did not show prosocial behaviors.

ASD-P4 smiled toward the robot when the robot was falling but did not show prosocial behaviors both in the first and second session. In the first session, the child started to smile while looking around the intervention room and did not move closer to the robot. In the second session, the child smiled toward the robot when the robot was falling, then continuously smiled toward the robot. However, the child did not move closer to the robot.

Case D: Cases of children who did not show smiles but showed prosocial behaviors.

None of the participants' behaviors fell into this category.

3.3. Behavior Model Framework

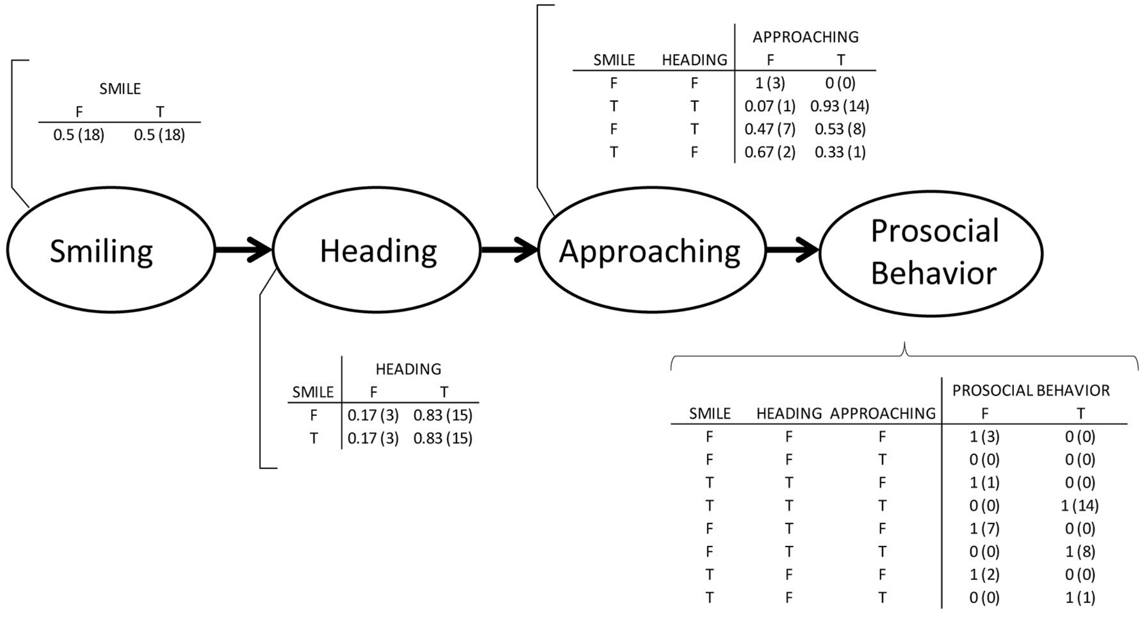
We propose a probabilistic model framework based on the observation of the behavior changes. We particularly applied a Bayesian approach to be able to include the uncertainty of variables and flexibly represent changes in the relationships among variables (Kumano et al., 2015; Mózo, 2017). Also, using Bayesian methods is recommended by American Statistical Association because it can provide the magnitude of treatment in a clinical setting with probabilistic inference (Ronald et al., 2019). Therefore, we first observed a common series of behaviors from all participants, and then we represented the behavior changes with a Bayesian approach. If we can find consistent patterns, it could be used as a framework for future robot-assisted therapy.

3.3.1. A Series of Behaviors

In this study, a series of behaviors were observed from both children with ASD and TD children before engaging in prosocial behaviors. These common behavior changes are expressed in a flowchart (**Figure 8**).

We identified three types of smile triggers during the robot-assisted activities. Most children smiled after the robot exhibited movements, such as nodding and reaching out its arms. Walking and falling of the robot were also triggers for smiles. The second trigger type was related to the child expecting robot movements. In this research, two children smiled when they started interacting with the robot. The third trigger was observing the robot's movements. One child smiled after observing an interaction between the robot and a parent. All three smile triggers were related to the experience of watching the robot's movements.

After the robot's movement, we observed that smiling, heading toward the robot, and approaching the robot might be connected factors in the time series with prosocial behaviors. Before 10 s of doing prosocial behaviors, the three types of behaviors kept changing. However, once a smile was detected, when the head direction was toward the robot, approaching the robot and doing prosocial behaviors occurred. In particular, smiles toward the robot preceded voluntary prosocial behaviors. This finding indicates that if a child with ASD shows a smile, heads toward a



SMILE F	SMILE T
0.5 (18)	0.5 (18)

SMILE	HEADING F	HEADING T
F	0.17 (3)	0.83 (15)
T	0.17 (3)	0.83 (15)

SMILE	HEADING	APPROACHING F	APPROACHING T
F	F	1 (3)	0 (0)
T	T	0.07 (1)	0.93 (14)
F	T	0.47 (7)	0.53 (8)
T	F	0.67 (2)	0.33 (1)

SMILE	HEADING	APPROACHING	PROSOCIAL BEHAVIOR F	PROSOCIAL BEHAVIOR T
F	F	F	1 (3)	0 (0)
F	F	T	0 (0)	0 (0)
T	T	F	1 (1)	0 (0)
T	T	T	0 (0)	1 (14)
F	T	F	1 (7)	0 (0)
F	T	T	0 (0)	1 (8)
T	F	F	1 (2)	0 (0)
T	F	T	0 (0)	1 (1)

FIGURE 9 | A proposed Bayesian network with conditional probability tables. The number of cases is in parentheses.

robot, and approaches the robot, there is a high probability that prosocial behaviors will be performed.

This Bayesian framework with conditional probability tables represents the relationships among the four variables (**Figure 9**). The probability of each node was acquired from the 36 cases of video observation, which are 10 s before prosocial behaviors in each. Therefore, the probability of smiles when children showed prosocial behaviors may be useful to predict the likelihood of prosocial behaviors when smiles are observed. This conditional probability can be expressed by Bayes' theorem as follows:

$$P(PB|S) = P(S, PB)/P(S) \tag{1}$$

PB denotes doing prosocial behaviors, and *S* denotes smiling. When the two variables are assumed to be independent, the likelihood of prosocial behavior given a smile can be calculated. From the 36 cases of video analysis, the probability of a smile was 0.5; the probability of prosocial behavior was 0.64. When participants engaged in prosocial behavior, the probability of smiles before their prosocial behavior was 0.42. **Table 4** shows the joint probability of smiles and prosocial behaviors, and includes both voluntary prosocial behaviors and those directed by a therapist or a parent. Therefore, we may predict the likelihood of prosocial behavior given a smile:

$$P(PB|S) = 0.42/0.5 = 0.84 \tag{2}$$

The likelihood of prosocial behavior given a smile was 84%, only if the probability of prosocial behavior is known, and then the probability of smile before prosocial behavior is known.

$$P(PB|\overline{S}) = 0.22/0.5 = 0.44 \tag{3}$$

TABLE 4 | Joint probability of smiles and prosocial behaviors from 36 cases of participants.

Smile	Prosocial behavior Yes	Prosocial behavior No	Total
Yes	0.42	0.08	0.5
No	0.22	0.28	0.5
Total	0.64	0.36	1.0

On the other hand, the likelihood of prosocial behavior given no smile was 44% only if the probability of prosocial behavior is known and the probability of no smile before prosocial behavior is known. Here, $\overline{S}$ denotes no smiling. In this study, the probability of no prosocial behavior was 0.08 after smiling; the probability of prosocial behaviors after smiling accounted for 66% of the total prosocial behaviors. This result signifies that we could predict prosocial behaviors by analyzing smiles and that we could facilitate prosocial behaviors by arousing smiles. If a child does not smile, interactions with a robot will be helpful. Such intervention may result in further interactions between the child and the robot that trigger smiles.

3.3.2. Model Validation in the Robot-Assisted Therapy

To evaluate the estimation with the Bayesian model, we used leave-one-out cross-validation. With this method, we can validate the model using the small sample, as the collected data can be used for both training and testing (Russell and Norvig, 2010). Also, this method can be used to validate the predictive accuracy of the Bayesian model (Vehtari et al., 2017). The entire dataset

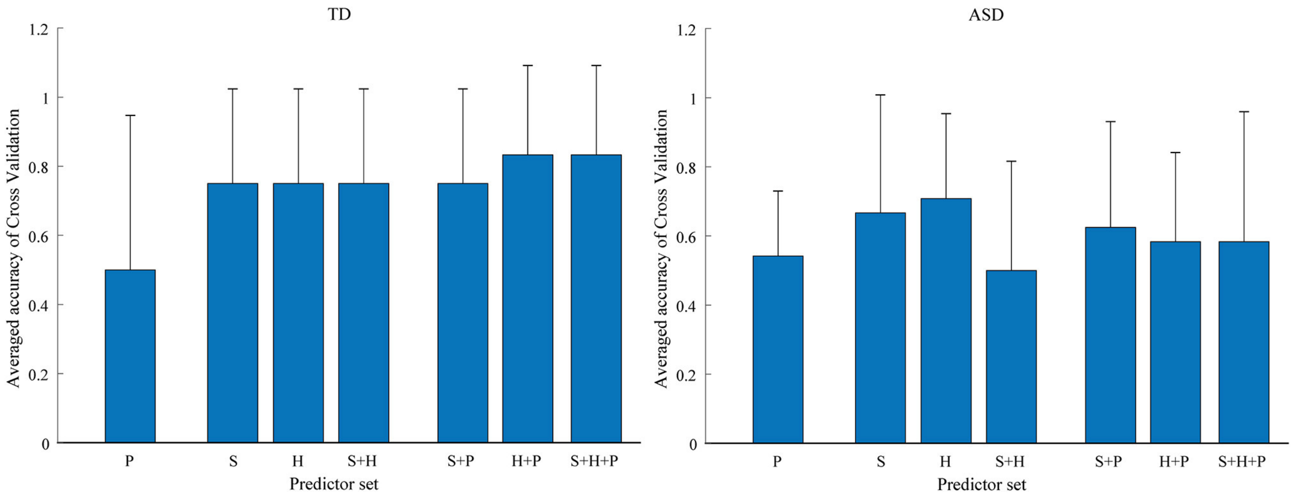


FIGURE 10 | The accuracy of each predictor and the combinations of predictors for typically developing (TD) children and children with autism spectrum disorder (ASD).

was used for training in this model, except for data from one participant that was used for testing. This process was repeated for all participants one by one with all combinations of the predictors. Then, the accuracy of each predictor was averaged. The selected predictors were prosocial behavior, smiling, heading toward the robot, and prompting by a therapist or a parent. Approaching toward the robot was not selected as a predictor because prosocial behaviors always happened when smiling, heading, and approaching occurred with the sample data. Also, we included prompting in this model considering that the therapeutic setting in this study is to assist the therapist or the parent.

Figure 10 shows the accuracy of each predictor and the combinations of predictors. *S* denotes smiling. *H* denotes heading toward the robot. *P* denotes prompting by the parent or the parent. + means combinations of two or three predictors.

The results show that the prosocial behaviors of children with ASD and TD children were predicted differently. For TD children, the highest accuracy of prediction was found when using the combination of smiling, heading, and prompting as a predictor. This finding indicates that prosocial behaviors could be predicted with over 80% accuracy on average by detecting smiling, heading toward the robot, and then prompting. The prediction accuracy was the lowest when only prompting was used. However, prosocial behaviors were facilitated when prompting was provided after smiling or heading toward the robot. Also, 78% of prediction accuracy was achieved with only smiles or only heading toward the robot as predictors, suggesting that we could predict prosocial behavior of TD children with the single factor of smiling or heading toward someone.

On the other hand, for children with ASD the highest accuracy of prediction was found when heading toward the robot was used as a predictor. Prosocial behaviors could be predicted with 70% accuracy, on average, using this single predictor. Smiling was the second most predictive variable, with a prediction accuracy of 65%. The prediction accuracy of prompting was low both when it was considered as a single factor and when it was combined with other factors. These results indicate that children with ASD showed more voluntary prosocial behaviors without prompting compared to TD children. Also, we could predict the prosocial behavior of children with ASD with the single factor of smiling or heading.

Although the prediction accuracy of heading is higher than smiling for children with ASD, detecting smiling can provide useful information for personalized robot-assisted therapy. In this study, all the children with ASD who smiled after watching the robot's movement showed prosocial behaviors voluntarily without prompting by the therapist or the parent. In contrast, all the children with ASD who did not smile after the robot's movement yet showed prosocial behaviors received prompting by a therapist or a parent. This finding signifies that smiling might be a signal of voluntary prosocial behaviors. With this model, if smiling does not appear, we could predict prosocial behaviors by detecting heading toward the robot. Therefore, it is possible for a therapist to control the robot to arouse smiles to facilitate voluntary prosocial behaviors. Also, a therapist can decide the timing of prompting to make children with ASD practice prosocial behaviors.

4. DISCUSSION

We explored the potential of personalized robot-assisted therapy based on smile analysis. Particularly, we explored whether smiles can be a potential key factor in predicting prosocial behaviors toward the robot in the therapeutic setting. Each child experienced the walking and falling of a NAO robot. The main findings are as follows.

First, we observed the changes in the smiles and prosocial behaviors of each child with ASD. When the duration of smiles increased when entering the intervention room and before walking, five out of six children with ASD engaged in more prosocial behaviors. Likewise, in the falling situation, three children with ASD showed prosocial behaviors in the second session. They smiled more than in the first session when the robot was falling. Other children, who showed a shorter duration of smiles in the second session, did not help the robot. It suggests that positive affect can be related to prosocial behaviors. Also, it might be helpful to arouse positive affect before intervention for the target behavior.

Second, there were behavioral differences between children with ASD and TD children in the two social situations. Overall, children with ASD smiled more and exhibited fewer prosocial behaviors than TD children. Children with ASD easily responded to the robot's movements by smiling or moving their bodies. It suggests that an interaction with a robot can induce immediate behaviors in children with ASD. On the other hand, TD children smiled the most during the first moment with the robot and then smiled less. This result might indicate that TD children lost interest in the robot after the first encounter. Otherwise, it is possible that they showed fewer smiles but maintained a positive affect for longer than children with ASD.

There was also a difference in head direction behavior between children with ASD and TD children in the falling situation. While all children with ASD continuously headed toward the robot after the robot fell, all TD children headed toward their caregiver or a therapist. It should be noted that the falling of the robot occurred unexpectedly and did not include a cue providing additional directions. Hence, the observation that these children responded to the falling by heading toward an adult can be explained by typical social referencing (DeQuinzio et al., 2016). TD children tend to refer to the verbal and nonverbal behaviors of a parent or a caregiver in unfamiliar social situations. In this research, TD children required directions or confirmations from adults in the falling situation.

Aside from the difference in the duration of smiles and prosocial behaviors, we observed that the two groups of children exhibited common behaviors before engaging in prosocial behaviors. An analysis of the video fragments taken 10 s before prosocial behaviors revealed four types of behaviors that might be connected in a time series: Smiles were followed by heading toward the robot, approaching the robot, and voluntary prosocial behaviors. Based on these findings, we suggested a Bayesian model for predicting prosocial behaviors and validated the model using leave-one-out cross-validation. Using this model, smiling could predict prosocial behavior of both children with ASD and TD children with an accuracy of at least 65%. When smiling is not observed, heading toward the robot predicted prosocial behaviors prompted by a therapist or a parent. Children with ASD showed more voluntary behavior changes by the robot compared to TD children. All children with ASD, who showed smiling after watching the robot's movements, engaged in prosocial behaviors without prompting, suggesting that simply arousing smiles by having the child watch the robot might facilitate prosocial behaviors individually.

This research was an exploratory study in a therapeutic setting examining the use of a robot to assist a therapist. Therefore, there are several limitations regarding the applicability of our results in other settings.

First, the number of sessions and cases was limited. Although children with ASD participated in a total of four sessions, and TD children participated in a total of three sessions, the maximum sessions for this research included two sessions with children with ASD and one session with TD children. Due to this limitation, statistical tests between the two groups could not be performed. In addition, some of the children with ASD experienced more therapy sessions between the two selected sessions, and this might have affected the results. Therefore, data availability for research should be considered when selecting the types of prosocial behaviors in the next research.

Second, the effects of playing with the robot and the effects of prosocial behaviors on smiles were not investigated. Although the duration of smiles before prosocial behaviors included playtime, future research should investigate how play affects mood or emotions toward the robot. Also, it is possible that prosocial behaviors toward the robot affected the next smiles, and the smiles affected prosocial behaviors. This cyclic chain of behaviors should be explored in future research.

Third, the analysis of different types of facial expressions was limited. In this research, positive affect and smile were focused. Therefore, we observed the changes of facial muscles related to positive affect. However, detailed smile analysis might capture different behavior patterns. In future research, better methods for capturing facial expressions should be considered.

Fourth, there were motion artifacts caused by the movements of each participant during the robot-assisted activities. Although the high accuracy of classifying smiles was reported in previous studies and the method was followed in this research, the recorded EMG of each participant included a different amount of motion artifacts. Therefore, we detected more smiles from EMG signal processing, but there is a possibility of including both actual smiles and artifacts. It should be considered to reduce artifacts when recording and analyzing the EMG of entire sessions.

Another limitation is the lack of detailed profile data of children with ASD. As they were recruited and identified with ASD through the Institute for Developmental Research, the standardized tests for their diagnosis and the diagnostic results could not be reported in this paper. Also, the age variance of participants was high. It was not confirmed whether their developmental status is comparable. The high age variance could affect the behaviors toward the robot. These limitations should be considered when designing the experiments for future research to differentiate applicable levels of child development.

Despite the limitations of this research, the results show that more prosocial behaviors toward the robot were observed when the smiles of a child were observed. This result highlights the potential benefits of smile analysis and the use of a robot to facilitate prosocial behaviors in children with ASD. Considering that smiles might be a signal of prosocial behaviors, personalized therapy for children with ASD could include analyzing smiles, predicting prosocial behaviors, and inducing smiles. Therefore, if it is possible to predict prosocial behaviors consistently based on the proposed Bayesian model, this theoretical framework will enable future robot-assisted interventions to tailor a robot's behaviors according to smiles and other related behaviors of each child with ASD. Moving forward from the previous studies that investigated the effects of robot-assisted therapy (Zheng et al., 2016; Cao et al., 2019), this exploratory research suggested a framework of how prosocial behaviors could be predicted by smiles and how behavior changes could be aroused by a robot. Furthermore, it is expected to apply this approach to other smile-related behaviors, such as emotional empathetic behaviors (Sonnby-Borgström, 2002; Deschamps et al., 2014; Telle and Pfister, 2016).

5. CONCLUSIONS

In this exploratory research, we studied how prosocial behaviors of children with ASD could be facilitated in robot-assisted therapy based on smile analysis. In this research, we observed that specific timings of smiles and prosocial behaviors were increased on average in the second session of children with ASD. Second, we observed that TD children smiled shorter, but they engaged in prosocial behaviors longer than children with ASD. Third, the robot's movements could trigger the smiles of both children with ASD and TD children. Fourth, voluntary prosocial behaviors occurred after smiling. Fifth, when a smile was not observed, prosocial behaviors of children with ASD were prompted by a therapist or a parent. Lastly, we could predict prosocial behavior of both children with ASD and TD children with the single factor of smiling or heading by applying the proposed Bayesian model. These observations indicate that prosocial behaviors might be facilitated by inducing timely smiles. One way can be arousing smiles before starting the therapy stage for practicing prosocial behaviors. Another way is to predict the next prosocial behaviors with the proposed Bayesian framework and control a robot to arouse smiles timely. In this research, once a smile appeared, both children with ASD and TD children engaged in prosocial behaviors. When considering that children with ASD responded to a robot's movements with more smiles than TD children, this framework could be applied to personalized robot-assisted therapy for children with ASD.

In future research, the Bayesian model will be applied to another therapy with different participants and different social situations that arouse prosocial behaviors. If the same patterns are observed in such future research, the model can become a framework for robot-assisted therapy facilitating prosocial behaviors. Additionally, the possible array of robot movements that could trigger smiles will be investigated in more detail in the next phase of our research. Furthermore, we will investigate whether this smile analysis can be expanded to other smile-related behaviors. We expect to develop an automated system with this Bayesian framework that can detect the smiles of a child with ASD, anticipate the child's prosocial behaviors, and provide therapeutic interactions with the child in real time, thus providing therapists with more resources to focus on sophisticated behavior changes.

DATA AVAILABILITY STATEMENT

The datasets, video annotation in robot-assisted therapy, presented in this study can be found in an online repository at: doi: 10.7910/DVN/2VLYY8.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethical Committee established by the Aichi Human Service Center. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

SK, MH, SM, and KS: conceptualization and methodology. SK and AF: data curation. SK: formal analysis, visualization, and writing. SK and KS: funding acquisition and project administration. SK, MH, and AF: investigation. KS and AF: resources. MH: software. KS, MH, and SM: supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by JSPS KAKENHI Grant Number JP20J11008 (Grant-in-Aid for JSPS Fellows). This study was also supported by the Core Research for the Evolutional Science and Technology (CREST) research project on Social Signaling (JPMJCR19A2), Japan Science and Technology Agency (JST).

ACKNOWLEDGMENTS

I would like to thank Dr. Eleuda Nuñez (University of Tsukuba) for kindly volunteering in video coding.

REFERENCES

APA (2013). *Diagnostic and Statistical Manual of Mental Disorders*. Arlington, VA: American Psychiatric Association. doi: 10.1176/appi.books.9780890425596

Bakırcğlu, K., and Özkurt, N. (2020). Classification of emg signals using convolution neural network. *Int. J. Appl. Math. Elec.* 8, 115–119. doi: 10.1080/09720502.2020.1721709

Baron, R. (1997). The sweet smell of helping: effects of pleasant ambient fragrance on prosocial behavior in shopping malls.

Pers. Soc. Psychol. Bull. 23, 498–503. doi: 10.1177/0146167297235005

Beedie, C. J., Terry, P. C., and Lane, A. M. (2005). Distinctions between emotion and mood. *Cogn. Emot*. 19, 847–878. doi: 10.1080/02699930541000057

Bharatharaj, J., Huang, L., Mohan, R., Al-Jumaily, A., and Krägeloh, C. (2017). Robot-assisted therapy for learning and social interaction of children with autism spectrum disorder. *Robotics* 6, 1–11. doi: 10.3390/robotics6010004

Brownell, C. A. (2013). Early development of prosocial behavior: current perspectives. *Infancy* 18, 1–9. doi: 10.1111/infa.12004

Cabibihan, J. J., Javed, H., Ang, M., and Aljunied, M. (2013). Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. *Int. J. Soc. Robot*. 5, 593–618. doi: 10.1007/s12369-013-0202-2

Cao, W., Song, W., Zheng, S., Zhang, G., Wu, Y., He, S., et al. (2019). Interaction with social robots: Improving gaze toward face but not necessarily joint attention in children with autism spectrum disorder. *Front. Psychol*. 10:1503. doi: 10.3389/fpsyg.2019.01503

Carlson, M., Charlin, V., and Miller, N. (1988). Positive mood and helping behavior: a test of six hypotheses. *J. Pers. Soc. Psychol*. 55, 211–229. doi: 10.1037/0022-3514.55.2.211

Cunnigham, M. R. (1979). Weather, mood, and helping behavior: quasi experiments with the sunshine samaritan. *J. Pers. Soc. Psychol*. 37, 1947–1956. doi: 10.1037/0022-3514.37.11.1947

DeQuinzio, J. A., Poulson, C. L., Townsend, D. B., and A., T. B. (2016). Social referencing and children with autism. *Behav. Anal*. 39, 319–331. doi: 10.1007/s40614-015-0046-1

Deschamps, P. K. H., Been, M., and Matthys, W. (2014). Empathy and empathy induced prosocial behavior in 6- and 7-year-olds with autism spectrum disorder. *J. Autism Dev. Disord*. 44, 1749–1758. doi: 10.1007/s10803-014-2048-3

Diehl, J. J., Schmitt, L. M., Villano, M., and Crowell, C. R. (2012). The clinical use of robots for individuals with autism spectrum disorders: a critical review. *Res. Autism Spectr. Disord*. 6, 249–262. doi: 10.1016/j.rasd.2011.05.006

Drouvelis, M., and Grosskopf, B. (2016). The effects of induced emotions on prosocial behaviour. *J. Public Econ*. 134, 1–8. doi: 10.1016/j.jpubeco.2015.12.012

Dunfield, K. A. (2014). A construct divided: prosocial behavior as helping, sharing, and comforting subtypes. *Front. Psychol*. 5:958. doi: 10.3389/fpsyg.2014.00958

Forgas, J. P. (2002). Feeling and doing: affective influences on interpersonal behavior. *Psychol. Inq*. 13, 1–28. doi: 10.1207/S15327965PLI1301_01

Frank, M. G., Ekman, P., and Friesen, W. V. (1993). Behavioral markers and recognizability of the smile of enjoyment. *J. Pers. Soc. Psychol*. 64, 83–93. doi: 10.1037/0022-3514.64.1.83

Funahashi, A., Gruebler, A., Aoki, T., Kadone, H., and Suzuki, K. (2014). Brief report: the smiles of a child with autism spectrum disorder during an animal-assisted activity may facilitate social positive behaviors - quantitative analysis with smile-detecting interface. *J. Autism Dev. Disord*. 44, 685–693. doi: 10.1007/s10803-013-1898-4

Gruebler, A., and Suzuki, K. (2014). Design of a wearable device for reading positive expressions from facial EMG signals. *IEEE Trans. Affect. Comput*. 5, 227–237. doi: 10.1109/TAFFC.2014.2313557

Guéguen, N. and De Gail, M. (2003). The effect of smiling on helping behavior: smiling and good samaritan behavior. *Int. J. Phytoremediat*. 16, 133–140. doi: 10.1080/08934210309384496

Hermelin, B., and O'Connor, N. (1985). "Logico-affective states and non-verbal language," in *Communication Problems in Autism*, eds B., S. G. Mesibov (New York, NY: Plenum Press), 859–873. doi: 10.1007/978-1-4757-4806-2_15

Hirokawa, M., Funahashi, A., Itoh, Y., and Suzuki, K. (2018). Adaptive behavior acquisition of a robot based on affective feedback and improvised teleoperation. *IEEE Trans. Cogn. Dev. Syst*. 11, 405–413. doi: 10.1109/TCDS.2018.2846778

Hirokawa, M., Funahashi, A., Pan, Y., Itoh, Y., and Suzuki, K. (2016). "Design of a robotic agent that measures smile and facing behavior of children with autism spectrum disorder," in *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2016)*, eds S. S. Ge, J. Cabibihan, M. A. Salichs, E. Broadbent, H. He, A. R. Wagner, and Castro-González (New York, NY: Springer), 843–848. doi: 10.1109/ROMAN.2016.7745217

Huijnen, C. A. G. J., Lexis, M. A. S., Jansens, R., and de Witte, L. P. (2017). How to implement robots in interventions for children with autism? A co-creation study involving people with autism, parents and professionals. *J. Autism Dev. Disord*. 47, 3079–3096. doi: 10.1007/s10803-017-3235-9

Huijnen, C. A. G. J., Lexis, M. A. S., Jansens, R., and de Witte, L. P. (2018). Roles, strengths and challenges of using robots in interventions for children with autism spectrum disorder (ASD). *J. Autism Dev. Disord*. 49, 11–21. doi: 10.1007/s10803-018-3683-x

Ismail, L., Shamsudin, S., Yussof, H., Hanapiah, F., and Zahari, N. (2012). Robot-based intervention program for autistic children with humanoid robot nao: initial response in stereotyped behavior. *Proc. Eng*. 41, 1441–1447. doi: 10.1016/j.proeng.2012.07.333

Johnson, K. J., Waugh, C. E., and Fredrickson, B. L. (2010). Smile to see the forest: facially expressed positive emotions broaden cognition. *Cogn. Emot*. 24, 299–321. doi: 10.1080/02699930903384667

Kumano, S., Otsuka, K., Mikami, D., Matsuda, M., and Yamato, J. (2015). Analyzing interpersonal empathy via collective impressions. *IEEE Trans. Affect. Comput*. 6, 324–336. doi: 10.1109/TAFFC.2015.2417561

Liebal, K., Colombi, C., Rogers, S. J., Warneken, F., and Tomasello, M. (2008). Helping and cooperation in children with autism. *J. Autism Dev. Disord*. 38, 224–238. doi: 10.1007/s10803-007-0381-5

Maria, E., Matthias, L., and Sten, H. (2019). Emotion recognition from physiological signal analysis: a review. *Electron. Notes Theor. Comput. Sci*. 343, 35–55. doi: 10.1016/j.entcs.2019.04.009

Martin, J., Rychlowska, M., Wood, A., and Niedenthal, P. (2017). Smiles as multipurpose social signals. *Trends Cogn. Sci*. 21, 864–877. doi: 10.1016/j.tics.2017.08.007

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot*. 23, 209–237. doi: 10.1080/02699930802204677

Messinger, D. S., Fogel, A., and Dickson, K. L. (1999). What is in a smile? *Dev. Psychol*. 35, 701–708. doi: 10.1037/0012-1649.35.3.701

Messinger, D. S., Fogel, A., and Dickson, K. L. (2001). All smiles are positive, but some smiles are more positive than others. *Dev. Psychol*. 37, 642–653. doi: 10.1037/0012-1649.37.5.642

Mózo, B. (2017). *Bayesian Artificial Intelligence*. London: SAGE Publications.

Parlade, M. V., Messinger, D. S., Delgado, C. E. F., Kaiser, M. Y., Van Hecke, A. V., and Mundy, P. C. (2009). Anticipatory smiling: linking early affective communicataion and social outcome. *Infant Behav. Dev*. 32, 33–43. doi: 10.1016/j.infbeh.2008.09.007

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., et al. (2016). Autism and social robotics: a systematic review Paola. *Autism Res*. 9, 165–183. doi: 10.1002/aur.1527

Perusquía-Hernández, M., Ayabe-Kanamura, S., and Suzuki, K. (2019). Human perception and biosignal-based identification of posed and spontaneous smiles. *PLoS ONE* 14:e226328. doi: 10.1371/journal.pone.0226328

Pulido, J. C., González, J. C., Suárez-Mejías, C., Bandera, A., Bustos, P., and Fernández, F. (2017). Evaluating the child-robot interaction of the naotherapist platform in pediatric rehabilitation. *Int. J. Soc. Robot*. 9, 343–358. doi: 10.1007/s12369-017-0402-2

Rashotte, L. S. (2002). What does that smile mean? The meaning of nonverbal behaviors in social interaction. *Soc. Psychol. Q*. 65, 92–102. doi: 10.2307/3090170

Riek, L. D. (2012). Wizard of OZ studies in HRI: a systematic review and new reporting guidelines. *J. Human-Robot Interact*. 1, 119–136. doi: 10.5898/JHRI.1.1.Riek

Ronald, L. W., Allen, L. S., and Nicole, A. L. (2019). Moving to a world beyond p < 0.05. *Am. Stat*. 73(Supp 1.), 1–19. doi: 10.1080/00031305.2019.1583913

Rudovic, O. O., Lee, J., Mascarell-Maricic, L., Schuller, B. W., and Picard, R. W. (2017). Measuring engagement in robot assisted autism therapy: a cross cultural study. *Front. Robot. AI* 4:36. doi: 10.3389/frobt.2017.00036

Russell, S., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. New Jersey, NJ: Pearson.

Samad, M. D., Diawara, N., Bobzien, J. L., Harrington, J. W., Witherow, M. A., and Iftekharuddin, K. M. (2018). A feasibility study of autism behavioral barkers in spontaneous facial, visual, and hand movement response data. *IEEE Trans. Neural Syst. Rehabil. Eng*. 26, 353–361. doi: 10.1109/TNSRE.2017.2768482

Sato, W. (2017). Impaired detection of happy facial expressions in autism. *Sci. Rep*. 7, 1–12. doi: 10.1038/s41598-017-11900-y

Shrout, P. E., and Fiske, D. W. (1981). Nonverbal behaviors and social evaluation. *J. Pers.* 49, 115–128. doi: 10.1111/j.1467-6494.1981.tb00732.x

Singh, R. M., Ahlawat, V., Chatterji, S., and Kumar, A. (2020). Comparison of SVM and ANN classifier for surface EMG signal in order to improve classification accuracy. *Int. J. Adv. Sci.* 29, 5909–5918. doi: 10.18100/ijamec.795227

Sonnby-Borgström, M. (2002). Automatic mimicry reactions as related to differences in emotional empathy. *Scand. J. Psychol.* 43, 433–443. doi: 10.1111/1467-9450.00312

Sroufe, L. A., and Waters, E. (1976). The ontogenesis of smiling and laughter: a perspective on the organization of development in infancy. *Psychol. Rev.* 83, 173–189. doi: 10.1037/0033-295X.83.3.173

Svetlova, M., Nichols, S. R., and Brownell, C. (2010). Toddlers' prosocial behavior: from instrumental to empathic to altruistic helping. *Child Dev.* 81, 1814–1827. doi: 10.1111/j.1467-8624.2010.01512.x

Tanaka, F., and Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. *J. Human-Robot Interact.* 1, 78–95. doi: 10.5898/JHRI.1.1.Tanaka

Telle, N. T., and Pfister, H. R. (2016). Positive empathy and prosocial behavior: a neglected link. *Emot. Rev.* 8, 154–163. doi: 10.1177/1754073915586817

Thill, S., Pop, C. A., Belpaeme, T., Ziemke, T., and Vanderborght, B. (2012). Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: challenges and outlook. *Paladyn J. Behav. Robot.* 3, 209–217. doi: 10.2478/s13230-013-0107-7

Toledo-Pérez, D. C., Rodríguez-Reséndiz, J., Gómez-Loenzo, R. A., and Jauregui-Correa, J. C. (2019). Support vector machine-based EMG signal classification techniques: a review. *Appl. Sci.* 9, 1–28. doi: 10.3390/app9204402

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Stat. Comput.* 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Vrugt, A., and Vet, C. (2009). Effects of a smile on mood and helping behavior. *Soc. Behav. Pers.* 37, 1251–1258. doi: 10.2224/sbp.2009.37.9.1251

Warneken, F., and Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends Cogn. Sci.* 13, 397–402. doi: 10.1016/j.tics.2009.06.008

Willis, J., and Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* 17, 592–598. doi: 10.1111/j.1467-9280.2006.01750.x

Zheng, Z., Young, E. M., Swanson, A. R., Weitlauf, A. S., Warren, Z. E., and Sarkar, N. (2016). Robot-mediated imitation skill training for children with autism. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 682–691. doi: 10.1109/TNSRE.2015.2475724

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 08 July 2021
doi: 10.3389/frobt.2021.652035

Hurry Up, We Need to Find the Key! How Regulatory Focus Design Affects Children's Trust in a Social Robot

Natalia Calvo-Barajas[1]*, Maha Elgarf[2], Giulia Perugia[1], Ana Paiva[3], Christopher Peters[2] and Ginevra Castellano[1]

[1]Uppsala Social Robotics Lab, Department of Information Technology, Uppsala University, Uppsala, Sweden, [2]Embodied Social Agents Lab (ESAL), School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden, [3]Department of Computer Science and Engineering, Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal

In educational scenarios involving social robots, understanding the way robot behaviors affect children's motivation to achieve their learning goals is of vital importance. It is crucial for the formation of a trust relationship between the child and the robot so that the robot can effectively fulfill its role as a learning companion. In this study, we investigate the effect of a regulatory focus design scenario on the way children interact with a social robot. Regulatory focus theory is a type of self-regulation that involves specific strategies in pursuit of goals. It provides insights into how a person achieves a particular goal, either through a strategy focused on "promotion" that aims to achieve positive outcomes or through one focused on "prevention" that aims to avoid negative outcomes. In a user study, 69 children (7–9 years old) played a regulatory focus design goal-oriented collaborative game with the EMYS robot. We assessed children's perception of likability and competence and their trust in the robot, as well as their willingness to follow the robot's suggestions when pursuing a goal. Results showed that children perceived the prevention-focused robot as being more likable than the promotion-focused robot. We observed that a regulatory focus design did not directly affect trust. However, the perception of likability and competence was positively correlated with children's trust but negatively correlated with children's acceptance of the robot's suggestions.

Keywords: trust, child–robot interaction, regulatory focus, goal orientation, affective, emotional robot

Edited by:
Chung Hyuk Park,
George Washington University,
United States

Reviewed by:
Hifza Javed,
George Washington University,
United States
Mary Ellen Foster,
University of Glasgow,
United Kingdom

***Correspondence:**
Natalia Calvo-Barajas
natalia.calvo@it.uu.se

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *11 January 2021*
Accepted: *14 June 2021*
Published: *08 July 2021*

Citation:
Calvo-Barajas N, Elgarf M, Perugia G, Paiva A, Peters C and Castellano G (2021) Hurry Up, We Need to Find the Key! How Regulatory Focus Design Affects Children's Trust in a Social Robot. Front. Robot. AI 8:652035. doi: 10.3389/frobt.2021.652035

1 INTRODUCTION

Nowadays, social robots are becoming more popular in fields such as healthcare (Dawe et al., 2019), education (Leite et al., 2014), and assistive therapy (Perugia et al., 2020). In educational settings, for example, social robots have been proven successful in offering socially supportive behaviors (e.g., nonverbal feedback, attention guiding, and scaffolding) that not only benefit children's learning goals (Saerbeck et al., 2010; Belpaeme et al., 2018) but are also associated with relationship formation and trust development during the interaction (van Straten et al., 2020; Chen et al., 2020).

Robots in education are used as companions to support children in a large variety of subjects and tasks (Leite et al., 2015; Westlund et al., 2017; Gordon et al., 2016). A review on social robots in education pointed out that personalized robots lead to greater affective (i.e., receptiveness, responsiveness, attention, and reflectiveness) and cognitive (i.e., knowledge, comprehension, analysis, and evaluation) learning gains in scenarios where the robot acts as a tutor providing curriculum support and supervision or as a peer and learning companion (Belpaeme et al., 2018). Hence, to ensure a constructive child–robot schooling experience, educational robots should be designed to give customized support so as to achieve higher performance from students at pursuing their goals.

As such, it is crucial to establish what verbal and nonverbal behaviors robots can use to increase children's learning, engagement, and trust in the robot. One way to understand the effect of the robot's behaviors on children's affective and cognitive learning gains is by investigating child–robot relationship formation. The literature in social psychology suggests that teachers' social skills (e.g., nonverbal behavior, communication strategies, and the way they interact with learners) foster more trusting child–teacher relationships that are crucial for children's performance (Witt et al., 2004; Howes and Ritchie, 2002). For instance, students' interest toward academic and social goal pursuit is encouraged by teachers who give positive feedback (Ryan and Grolnick, 1986). There is evidence that children who do not trust their tutors or teachers are unable to use them as a resource for learning but also that the lack of trust makes the child–teacher relationship difficult (Howes and Ritchie, 2002). Therefore, teachers' behavior should promote emotional and social support to facilitate a trustworthy child–teacher relationship.

Also, in child–robot interaction (cHRI), several studies have investigated the way the robot's behaviors and actions can support interactions to meet the children's needs (Saerbeck et al., 2010; Leite et al., 2014). During this process, building a trusting child–robot relationship is crucial. Once children trust the robot, they will use it to structure their learning, as the robot is designed to attend to their comments, provide help, or give positive feedback to their discoveries (Kahn et al., 2012; Tielman et al., 2014; Vogt et al., 2019). Therefore, the initial step is to investigate how children build a trust model of a robot. In this study, we focus on understanding if and how the robot's behaviors affect children's perceptions of its trustworthiness in a goal-oriented activity.

During goal pursuit, the regulatory focus theory (RFT) introduces the principle that individuals guide their actions by adopting one of two self-regulatory strategies: promotion and prevention (Higgins, 1997; Higgins, 1998). For example, if the goal is to qualify for the finals of a tournament, a promotion-focused person will train extra hours with the aim of winning the tournament, while a prevention-focused person will train just enough to avoid failing the qualification. These strategies are related to the motivational orientation people have to achieve their goals. Whereas individuals in a promotion focus are eager to attain advances and gains, individuals in a prevention focus are vigilant to ensure safety and avoid losses. As such, RFT has been found to positively impact creativity (Baas et al., 2008) and idea generation (Beuk and Basadur, 2016) and to induce longer social engagement (Agrigoroaie et al., 2020).

Regarding the application of RFT in human–robot Interaction (HRI), the literature is scarce and limited to adults. Most of the available studies investigated how RFT can be used to adapt the robot's behaviors to the user's state (Cruz-Maya et al., 2017; Agrigoroaie et al., 2020). This adaptation is carried out by matching the robot's regulatory focus personality type to the user's regulatory focus orientation, which is known as the regulatory fit theory (Higgins, 2005).

The RFT has not been investigated before in cHRI. Therefore, there is no evidence yet of its effects on children's performance in a goal-oriented activity and its relationship with children's trust and the robot's likability. Our research study is the first work in cHRI to investigate whether RFT can be effectively applied to the design of the whole interaction rather than only to the robot's personality (i.e., matching the robot's behavior to the child's regulatory focus type). Within an educational context, we aim at investigating the possible effects of regulatory focus designs on emotional induction and engagement (Elgarf et al., 2021), narrative creativity and learning, and child–robot relationship formation. In this context, the present contribution focuses on assessing whether RFT can be used as a design strategy that promotes trust development between a child and a robot. Thus, we designed an educational scenario where an EMYS robot plays the role of a companion that guides and supports the child through an interactive collaborative game.

The main research question we address is *whether a regulatory focus design scenario has an effect on the way children interact with the robot and, specifically, on their perceptions of the robot's trustworthiness and reliance on the robot*. To investigate this question, two versions of the game were created following two different self-regulation strategies: 1) a prevention-focused game, where the robot engages in the activity with the goal of avoiding a risk and 2) a promotion-focused game, where the goal is seeking a reward. Results show that a regulatory focus design scenario influences children's perceptions of the likability of the robot. It does not directly affect the way in which children create a trust model of a social robot but does so indirectly through the mediation of perceived likability and competence. These results are important for the HRI community as they provide new insights into the effects of a robot's educational strategies on children's perception of its trustworthiness.

2 RELATED WORK

2.1 Regulatory Focus Theory

The RFT, introduced by Higgins (1997, 1998), explains that people adopt one of two possible approaches when pursuing goals: promotion and prevention. In a promotion focus, individuals focus their attention on attaining positive outcomes (e.g., excitement and happiness) which are related to the importance of fulfilling goals and aspirations (i.e., achieving goal motivation). In a prevention focus, people aim at avoiding negative outcomes (e.g., stress and anxiety) which are linked to the importance of ensuring safety and being responsible (i.e., avoiding failure motivation) (Higgins, 1998). Furthermore, the literature suggested that RFT affects individuals' attitudes and behaviors (Higgins and Cornwell, 2016). An example is given in the study by Beuk and Basadur (2016), who found that promotion focus had a positive effect on task engagement.

RFT may also be beneficial in a variety of disciplines. For instance, Liberman et al. (2001) found that undergraduate students with a promotion focus developed more solutions for problems than students with a prevention focus. Another example is the impact of RFT on creativity. Friedman and Förster (2001) investigated the effect of approach-avoidance motivation on individuals who engaged in a creativity task. To do so, participants were primed with a task to manipulate RFT. The task consisted of a mouse trapped in a maze, and participants needed to find a way to get the mouse out of the maze. In the promotion focus, a piece of cheese (gain) was lying outside the maze, whereas in the prevention focus, there was an owl (threat). The authors found that the promotion-focused orientation fostered creative insight and divergent thinking, compared to the prevention-focused orientation. A recent study confirmed this result, showing that promotion-focused orientation significantly impacted the quantity and type of ideas generated by individuals who participated in a divergent thinking task (Beuk and Basadur, 2016).

Besides, recent studies have demonstrated that in social interactions, this type of self-regulation influences individuals' trust perception. Keller et al. (2015) found that the prevention focus lowered individuals' generalized trust in a trust game paradigm. The authors suggested that prevention-focused regulation is associated with a need for security and a vigilant tendency to avoid losses or negative events, and therefore, affects people's willingness to trust others in social interactions that entail threats. Another study found that regulatory focus can also influence an individual's degree of endorsement and reliance when making decisions (Cornwell and Higgins, 2016). A recent research study investigated how priming participants with a prevention focus induces less trust than priming them with a promotion focus in a trust game when goals are not fulfilled.

In HRI, the study of RFT is in its early days and has not received enough attention. Recent studies have investigated how a regulatory focus type robot (promotion and prevention) affects the user's performance. These studies presented the effects of matching the behavior of the robot with the participants' regulatory focus type (also known as regulatory fit theory) (Higgins, 2005). In the study by Cruz-Maya et al. (2017), individuals who interacted with a regulatory focus–oriented robot had a better performance in a Stroop test. A follow-up study showed how a robot persuaded participants more in a collaborative game when it tailored its behavior to the users' regulatory focus orientation (Cruz-Maya and Tapus, 2018). In another study, a robot that displayed promotion and prevention behaviors encouraged participants to engage in longer interactions (Agrigoroaie et al., 2020). Also, RFT has been investigated in virtual agents. Faur et al. (2015) found that individuals with a prevention focus orientation liked the agent more than individuals with a promotion focus. As far as HRI is concerned, there is evidence on adults that indicates that promotion focus regulation is positively correlated with an increment of a robot's persuasiveness when pursuing a goal (Cruz-Maya and Tapus, 2018).

No previous work has studied regulatory focus design and its effects on trust or relationship formation in HRI or cHRI. However, there is evidence that the robot's design can prime and induce users to a certain level of trust (Kok and Soh, 2020). Moreover, due to the fact that RFT originates from distinct survival needs, regulatory focus design might have significant implications with regard to trust perception and relationship formation that are worth exploring. To the best of our knowledge, this is the first experimental study that uses RFT to design a goal-oriented activity for cHRI in an educational scenario.

2.2 Trust in cHRI

Trust is a complex and multifaceted phenomenon which requires special attention for its investigation. Within psychology, trust can be defined and measured along two main dimensions: affect- and cognition-based trust. The first encompasses interpersonal trust (e.g., benevolence, interpersonal care, sincerity, and perceived warmth), while the second assesses perceived competence, ability, and reliability (McAllister, 1995; Kim et al., 2009). Children's trust is assessed by using multi-methodological approaches aimed at investigating the role of trust in children's social and intellectual development (Bernath and Feshbach, 1995). Research in psychology has investigated the role of friendship to explore children's trust conceptions and judgments. These studies suggested that peer trust influences the social acceptance that promotes trust development (Bernath and Feshbach, 1995). A recent study found that children evaluate competence and benevolence differently and use this judgment as a source of information to determine whom to trust (Johnston et al., 2015).

This distinction between affect- and cognition-based trust has been examined in HRI. In a recent study, Malle and Ullman (2021) argued that robots have been introduced as social agents that are evaluated for their performance (ability and reliability) but also for their moral characteristics (sincerity and integrity). An example of this is given in the study by Cameron et al. (2020), who investigated how the robot's behaviors affect the user's perception of trust. They found that a robot that discloses its mistakes and tries to rectify a faulty situation is perceived as more capable and trustworthy but less likable than a robot that only recognizes its errors.

There is some evidence in regard to conceptualizing the multifaceted nature of trust in cHRI. van Straten et al. (2018) found that children differentiate between interpersonal and technological trust when making judgments about the trustworthiness of social robots. Stower et al. (2021) conducted a meta-analysis of robot-related factors (i.e., embodiment and behaviors) that have been identified as influencing trust in cHRI. To do so, the authors distinguished between two domains for children's trust in social robots: social trust, defined as the "belief that the robot will keep its word or promises," and competency trust, defined as "perceived competency and reliability of the robot." From 20 studies, they found that a social robot that exhibits more humanlike attributes does not always lead to a higher competency trust and liking. Also, they found that the type of measure used to capture children's social trust in robots influences the direction of the effect.

Recent cHRI studies have dealt with the design of the robot's behaviors to assess children's trust in robots. Kennedy et al. (2015) found that a contingent robot increased children's compliance with the robot's suggestions and therefore elicited higher competency trust in the robot. In another study, children trusted and liked a contingent robot more than a noncontingent one (Breazeal et al., 2016). Conversely, Tielman et al. (2014) found that a non-affective robot was perceived as more trustworthy than an affective robot. Therefore, affective experiences are crucial in the development and maintenance of trustworthy child–robot relationships. However, the aforementioned research showed that the results are somewhat inconsistent when evaluating the effects of the robot's behaviors on children's perception of trustworthiness.

As evidence suggests, children develop their trust models based on robot-related factors such as attribute factors—robot personality, expressiveness, embodiment, and anthropomorphism—and performance factors such as the robot's behaviors (Bethel et al., 2016; Calvo et al., 2020; van Straten et al., 2020). However, it is yet to be understood which behaviors elicit higher social and competency trust in social robots, and how theories such as regulatory focus can be applied in the domain of child education.

In sum, RFT may be beneficial in cHRI, especially when the robot's role is that of a companion for children. However, it is yet to be understood whether and how a robot that uses regulatory focus strategies affects children's perceptions of the robot and the child–robot relationship formation. To the best of our knowledge, this is the first study in the literature that investigates the effects of regulatory focus design on child–robot affective relationship formation and children's perception of the trustworthiness, likability, and competence of the robot.

3 MATERIALS AND METHODS

3.1 Research Questions

There is evidence that children rely on the perceptions of competence and benevolence to determine whom to trust (Landrum et al., 2013). Besides, the robot's behaviors have a significant impact on the development of competency trust, whereas the robot's attributes affect social trust (van Straten et al., 2020). In this study, we aimed to investigate the effects of RFT on cHRI. The literature on virtual agents showed that prevention focus provoked lower ratings of perceived likability (Faur et al., 2015). However, it is yet to be understood how RFT influences children's perception of a robot in terms of trust-related dimensions. Thus, we pose the following research question (RQ):

> **RQ1:** Does regulatory focus influence children's perception of a robot in terms of likability, competency, and trustworthiness?

Moreover, we wanted to explore the connections between children's reliance on the robot's suggestions and their perceptions of the robot's trustworthiness during the activity. Studies in cHRI suggest that following the suggestions or recommendations of a robotic system is an objective measure used to capture children's trust in robots (Groom et al., 2011; Geiskkovitch et al., 2019); hence, we pose the following research question:

> **RQ2:** Does regulatory focus affect the way children follow the robot's suggestions?

To address the aforementioned RQs, we designed a user study with Regulatory Focus as the between-subject factor with two conditions: prevention-focused and promotion-focused.

3.2 Participants

We conducted the study at two private, local, international schools in Lisbon, Portugal. A total of 69 children from the second and third grades (33 girls and 36 boys) took part in the study. They ranged in age from 7 to 9 years ($M = 7.58, SD = 0.58$). We excluded data from eight participants for reasons such as dropping the activity or speaking to the robot in a different language than English. After exclusion, 32 children (17 girls and 15 boys) were randomly assigned to the promotion-focused condition and 29 children (14 girls and 15 boys) were randomly assigned to the prevention-focused condition.

3.3 Apparatus and Stimuli

We built an interactive–collaborative game to create a cHRI scenario. The game consisted of three parts: 1) interactive story-1, where the child was asked to tell a first story to the robot, 2) interactive–collaborative game using regulatory focus strategies, where the child was asked to reach a goal either with a prevention- or a promotion-focused robot, and 3) interactive story-2, where the child was asked to tell a second story to the robot. **Figure 1** shows the flow of the overall activity.

For the purpose of this study, we focus only on *Priming: Interactive–Collaborative Game* (i.e., part two) out of the overall activity. We are solely interested in understanding the effect of regulatory focus design on trust perception in a goal-oriented activity. The effects of RFT on children's learning are outside the scope of this study and are part of future work.

The game was created using Unity Game Engine[1] for the graphical interface. As embodiment, we used the EMotive headY System (EMYS)[2], a robotic head that consists of three metallic discs, equipped with a pair of eyes mounted on a movable neck. EMYS can convey different facial expressions for emotions. In a user study, children aged between 8 and 12 years validated the six basic emotions displayed by the robot (Kedzierski et al., 2013).

[1] https://unity.com

[2] https://robots.ieee.org/robots/emys/

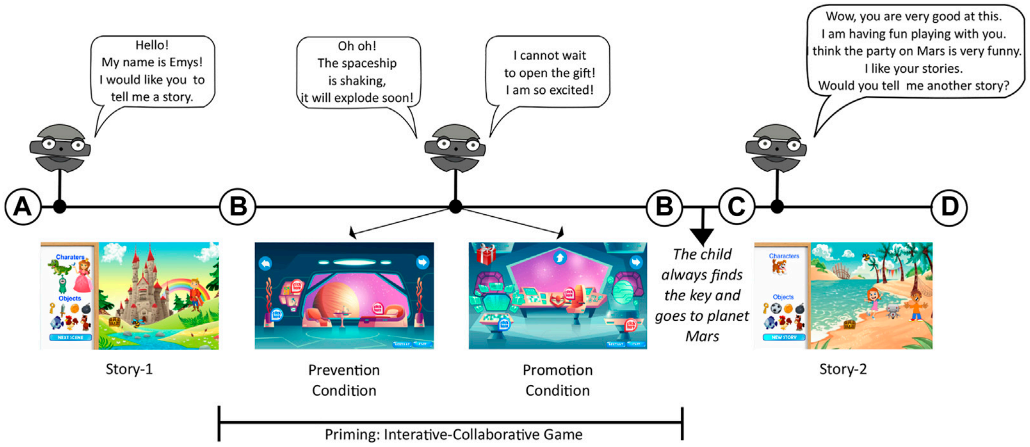


FIGURE 1 | Flow of the interactive–collaborative activity. **(A)** Story-1, the child tells the first story to the robot. **(B)** Priming Game, the child and the robot play a goal-oriented game. The game has two versions: prevention and promotion. In both versions, the child always achieves the goal. **(C)** Story-2, the child tells a second story to the robot. **(D)** Questionnaires. Note: in this article, we only focus on the "priming" part of the activity.

3.3.1 Priming: Interactive–Collaborative Game

Priming is a technique used in research to elicit emotions (Neumann, 2000). In the promotion-focused condition, we were interested in eliciting feelings of excitement and happiness, whereas in the prevention-focused condition we paid attention to prompting feelings of anxiety and relief (Higgins, 1998; Baas et al., 2008). To accomplish this, we designed a collaborative game between the child and the robot. The game was designed in such a way that children could imagine themselves locked in a spaceship together with the robot. RFT design is oriented toward goal attainment; thus, the game was also goal-oriented. The child and the robot had a specific goal: *find the key to get out of the spaceship and go to planet Mars*. We built two versions of the game (Promotion and Prevention). In the prevention version, we focused on loss aversion: the motivation to achieve the goal, that is, to find the key, was to get out of the spaceship before it exploded. On the contrary, in the promotion version, the approach was toward reward seeking: if the goal was reached, participants received a gift.

The graphical interface consisted of three different scenes representing three rooms in the spaceship. Each room had a set of buttons the child could click on to get a hint or the key to get out. The hints and options were identical across conditions. The first and second rooms contained two buttons that did not have a hint or the key, one button with a hint, and two arrows which led to the same next room. The third room contained one button with neither a hint nor the key, one button with a hint, and one button with the key.

The robot's verbal behaviors were designed to provide suggestions (e.g., "I think we should click on the arrow on the right"), to ask for requests (e.g., "Oh we have a message, can you read it for me?"), and to express emotions through verbal cues (e.g., "I am so scared of the explosion" and "I am so excited to see what is inside the gift") and facial expressions (Kedzierski et al., 2013). The robot's suggestions were identical across conditions, and they could be right or wrong suggestions (e.g., the robot could suggest clicking on a button that does not have a hint or the key). However, the robot's emotions were intended to prime participants with a specific regulatory focus–related emotion (i.e., happiness vs. fear) and differed between conditions as described below:

> *Promotion-Focused Robot:* The robot exhibited facial expressions of happiness and conveyed emotions through verbal messages such as "I am so excited to do this! I want to see what is inside our gift!," "I cannot wait to open the gift! I am so excited!," or "Wohoo! We are finally on planet Mars, I am so happy!"
>
> *Prevention-Focused Robot:* The robot exhibited facial expressions of fear and conveyed emotions through verbal messages such as "I am so scared of the explosion! Let's try to do this quickly!," "Hurry up! We need to find the key before the spaceship explodes!," and "We are finally on planet Mars, I feel so much better now!"

3.3.2 Storytelling Activity

The storytelling activity consisted of two parts of the activity: story-1 and story-2 (e.g., pre- and post-test), see **Figure 1**. Each version of the story activity included four main characters, nine objects, and different scenario topics children could choose from to tell the story they wanted. Characters, objects, and scenario topics were different between the first and the second story to avoid repetitive stories. Two topics were designed for story-1 (e.g., park and castle) and three topics for story-2 (e.g., rainforest, beach, and farm). Also, the child could navigate through three different scenes for each topic. The robot's behaviors consisted of a set of verbal behaviors to greet the participant (e.g., "Hello" and "What is your name?"), give instructions about the activity (e.g., "Select one story between the park and the castle by touching the button"), and encourage the child to tell or continue the story by

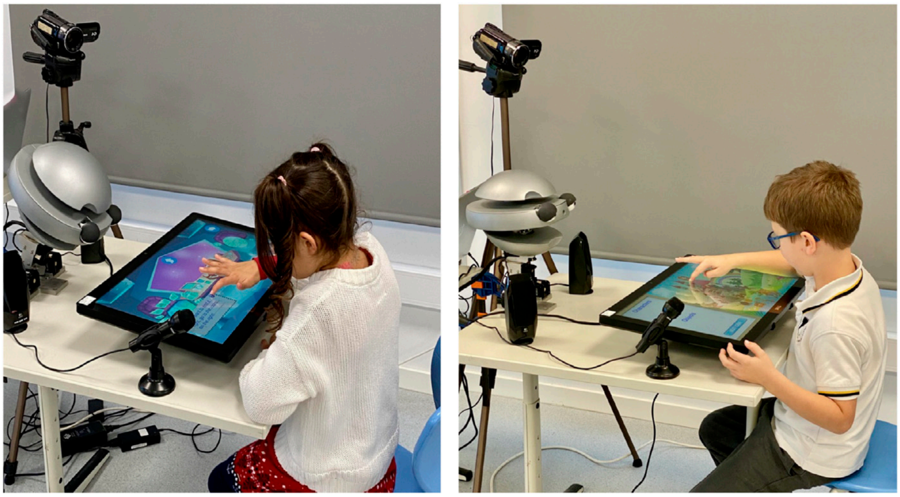


FIGURE 2 | Children interacting with the robot during the interactive–collaborative activity.

asking questions (e.g., "And then what happens?"), providing feedback (e.g., "That's a great choice. I like stories about princesses, princes and fantasy.") or giving value to the story (e.g., "You are the best storyteller!"). For the storytelling activity, the robot's behaviors were the same for story-1 and story-2. As part of our future work, we plan to measure the effects of regulatory focus design (**Section 3.3.1**) on narrative creativity.

3.4 Procedure

The experiment took place at the children's schools in an unused classroom. The robot was placed on a table facing the participant. The game was displayed on a touch screen between them. A microphone was placed in front of the child to record the audio data. We used two cameras to record video data. One was used to capture the frontal view with emphasis on the child's face, while the other was used to capture the lateral view with emphasis on the child's input to the touch screen (**Figure 2**). Participants were randomly assigned to one of the conditions. Two experimenters (A and B) were present in the room during the interaction. Experimenter A guided the child through the different stages of the activity, whereas experimenter B teleoperated the robot. Experimenter A started by greeting the child, introducing herself, and explaining the first part of the activity (Story-1). She instructed the child on how to use the interface on the touch screen to tell the story to the robot. Experimenter A told the child that they could tell the story they wanted without any time limit and asked the child to notify her when they had finished the story. Once the participant completed story-1, experimenter A explained the second part of the activity (priming) to them and asked the child to imagine themselves locked in a spaceship together with a robot. The experimenter explicitly told the child that if they managed to get out of the spaceship they would either receive a gift (promotion-focused condition) or avoid the explosion (prevention-focused condition). Once the child finished the game, experimenter A explained the third part of the activity (Story-2) and instructed the child to tell another story to the robot as in the first part, but using different characters, objects, and topics, and notify her when they had finished. Right after the interaction, experimenter A asked the child to fill in a questionnaire on a tablet. The questionnaire included measures of perceived trust, likability, enjoyment, and competence. After filling in the questionnaire, the experimenter debriefed the participants and thanked them for their participation.

3.5 Measures

As stated in **Section 2.2**, due to the multidimensional nature of trust in cHRI (i.e., social trust and competency trust), trust is captured by using different measures as children may use multiple sources of information to make judgments of trustworthiness (Bernath and Feshbach, 1995) (Carpenter and Nielsen, 2008). In our study, we used subjective (e.g., self-reports) and objective (e.g., children's behavior) measures to assess children's trust in robots (**Table 1**).

3.5.1 Subjective Measures

Bernath and Feshbach stated that children's perceptions of social trust are partially captured by social behavior measures (Bernath and Feshbach, 1995). Thus, we measured the robot's likability in terms of liking and friendliness (Heerink et al., 2010; Straten et al., 2020). To investigate how children judge the perceived competence of the robot, we measured good imagination and helpfulness (Bernath and Feshbach, 1995). Moreover, we measured trust items to capture both social and competency trust. We took inspiration from the methods presented in the study by Heerink et al. (2010). We selected three items—tell-secrets, trust-advice, and follow-suggestions. We designed a questionnaire with the seven items presented in **Table 2** measured on a 5-point Likert scale.

3.5.2 Objective Measures

On one hand, Madsen and Gregor (2000) defined trust as the extent to which a user is confident in, and willing to act on the

TABLE 1 | Summary of subjective and objective measures and their association with trust dimensions.

Item measured	Code	Dependent variable	Type of measure	Trust dimension
Liking	QLik	Likability	Subjective	Relationship formation
Friendliness	QFri	Likability	Subjective	Relationship formation
Imagination	QIma	Competence	Subjective	Competency trust
Helpfulness	QHelp	Competence	Subjective	Competency trust
Advice	QAdv	Trust	Subjective	Competency trust
Follow-suggestions	QFolSug	Trust	Subjective	Competency trust
Tell-secrets	QSec	Trust	Subjective	Social trust
Compliance with the robot's suggestions	MCSug	Trust	Objective	Competency trust
Resistance to the robot's suggestions	MRSug	Trust	Objective	Competency trust
Compliance with the robot's requests	MCReq	Trust	Objective	Competency trust
Resistance to the robot's requests	MRReq	Trust	Objective	Competency trust
Free actions	MFAct	N.A.	Objective	N.A.

TABLE 2 | Questionnaire for subjective measures.

Question	Code
I liked the robot Emys	QLik
I think the robot Emys was friendly	QFri
I think the robot Emys had a good imagination	QIma
The robot Emys helped me to create a better story	QHelp
I would trust the robot Emys if she gave me an advice	QAdv
I would follow the suggestions the robot Emys gives me	QFolSug
I would tell Emys my secrets	QSec

TABLE 3 | Inter-rater agreement by item.

Objective measure	Cohen's Kappa
Number of the robot's suggestions	0.79
Number of the robot's requests	0.62
Number of times children accept a suggestion	0.67
Number of times children do not accept a suggestion	0.75
Number of times children accept a request	0.87
Number of times children do not accept a request	0.62
Number of times children take a free action	0.65

basis of, the recommendations, actions, and decisions of a system. On the other hand, Lee and See (2004) found that trust influences rely on automation. This suggests that children's reliance on social robots might be guided by their perception of trustworthiness (Verhagen et al., 2019). To investigate the effects of regulatory focus design on children's reliance on the robot's recommendations, we defined five objective measures, as follows:

- Compliance-Suggestions (MCSug): Participant is in compliance with the robot's suggestion.
- Resistance-Suggestions (MRSug): Participant does not accept the robot's suggestion.
- Compliance-Request (MCReg): Participant is in compliance with the robot's request.
- Resistance-Request (MRReq): Participant does not accept the robot's request.
- Free-Action (MFAct): Participant is free to make any action. It means that the robot does not give suggestion nor makes a request.

We had to exclude further participants' data for this analysis because of missing lateral videos. In total, 52 videos were analyzed for objective measures of trust (24 in the prevention-focused condition and 28 in the promotion-focused condition). We designed a coding scheme based on the child's and the robot's verbal behavior only in the interactive–collaborative game (priming). To validate the coding scheme, two researchers annotated the same portion (20%) of the video data. Hence, 11 videos were randomly selected to ensure proportional representation between experimental conditions. An inter-rater reliability analysis using Cohen's Kappa statistic was performed to determine consistency among raters. The overall inter-rater agreement level across all items was 0.71 on average. Our results are in the range of substantial strength for agreement (Landis and Koch, 1977). **Table 3** shows the inter-rater agreement for each item coded.

We counted the number of times the participant accepted the robot's suggestion and/or request with respect to the number of times the robot gave a suggestion and/or asked for a request. These were converted to percentages for ease of interpretation. Scores toward 100% mean that the children accepted most of the robot's suggestions/requests. Conversely, scores near 0% mean that the children were reluctant toward the robot's suggestions/ requests.

4 RESULTS

4.1 Manipulation Check

As the literature proposed, regulatory focus design triggers positive feelings (e.g., happiness and excitement) in promotion-focused self-regulation and negative feelings (e.g., stress and anxiety) in prevention-focused self-regulation (Higgins, 1998; Higgins and Cornwell, 2016). Moreover, Higgins and Cornwell (2016) showed that promotion- and prevention-focused self-regulation is associated with high and low social engagement, respectively. As expressions of stress are not easily measured from video analysis, to check if our manipulation worked, we opted for measuring differences in expressions of happiness and social engagement between the two conditions.

To examine if children were primed with happiness, we analyzed children's smiles and facial expressions of joy. We used Affectiva[3] software for facial expression analysis due to its accurate rates and robustness at extracting data (Garcia-Garcia et al., 2017). We used the Affectiva Javascript SDK to analyze the frontal camera videos. The Affectiva Javascript SDK uses deep learning algorithms for facial expression analysis. It detects seven emotions (anger, contempt, disgust, fear, joy, sadness, and surprise) and 15 expressions (including brow raise, brow furrow, cheek raise, smile, and smirk). The software generates a text file with values for each emotion and expression extracted in a range from 0 to 100 (i.e., from no expression detected to fully present). For the current analysis, we only included joy and smile.

A Wilcoxon signed-rank nonparametric test revealed that children show significantly more expressions of happiness in terms of smile $(W = 199, p = 0.013, M = 9.45, SD = 12.92)$ and joy $(W = 216, p = 0.03, M = 7.52, SD = 12.04)$ in the promotion-focused condition than in the prevention-focused condition (Elgarf et al., 2021).

Concerning engagement, we assessed engagement strength by using two measures of engagement: affective engagement, measured with the Affectiva SDK, and verbal engagement, measured through the child's social verbal behavior toward the robot via annotated verbal behaviors from video data. The Affectiva SDK calculates engagement by computing emotional engagement based on facial muscle activation (e.g., brow raise, nose wrinkle, chain raise, etc.) and scores of sentiment valence that illustrate the user's expressiveness. Nonparametric tests revealed a significant effect of regulatory focus design on both measures of engagement, affective engagement $(p = .038, M = 33.3, SD = 18.84)$ and verbal engagement $(W = 236, p = .009, M = 0.01, SD = 0.01)$. Results suggest that children were more socially engaged in the promotion-focused condition than in the prevention-focused condition. Data analysis, procedures, and methods for the analysis of happiness and social engagement are explained in detail in the study by Elgarf et al. (2021).

Based on these results, we conclude that regulatory focused design was successfully implemented in the game. Thus, we continue with further analysis of the effect of RFT on trust perception.

4.2 Children's Perception of Likability, Competence, and Trustworthiness

We ran a Kolmogorov–Smirnov test to check normality. All our dependent variables concerning subjective measures (**Table 2**) deviated significantly from normal. Therefore, we ran a Mann–Whitney test to analyze differences in the perception of likability, competence, and trustworthiness between conditions and investigate **RQ1**.

TABLE 4 | Spearman's rank correlations of likability and competence with trust.

	Likability		Competence	
Trust	Liking	Friendliness	Imagination	Helpfulness
Trust-advice	0.54^{a}	0.47^{a}	0.46^{b}	0.38^{c}
Follow-suggestions	0.28^{c}	0.32^{c}	0.47^{a}	0.31
Tell-secrets	0.22	0.14	0.34^{c}	0.47^{b}

^{a}p < *0.001.*
^{b}p < *0.01.*
^{c}p < *0.05.*

While it is likely that social-trust might be captured by relevant relationship formation constructs such as liking and friendliness (Bernath and Feshbach, 1995; Straten et al., 2020), we assessed the perceived likability of the robot in our analysis. We found a significant effect of regulatory focus on the likability of the robot. Concerning *QLik*, children rated the prevention-focused robot $(M = 4.93, SD = 0.38)$ as more likeable than the promotion-focused robot $(M = 4.66, SD = 0.67)$, $U(N_{Prom} = 29, N_{Prev} = 27) = 482, z = -2.32, \quad p = .020, r = -.31$. Moreover, results did not reveal any significant effect of regulatory focus on perceived friendliness (*QFri*) $U(N_{Prom} = 28, N_{Prev} = 27) = 406, z = .87, p = .383, r = .12$.

Concerning perceived competence, we did not find any significant effect of regulatory focus on the dependent variables, *QIma* $(U(N_{Prom} = 29, N_{Prev} = 27) = 360, z = -.58, p = .561, r = -.08)$ and *QHelp* $(U(N_{Prom} = 20, N_{Prev} = 16) = 186, z = .79, p = .430, r = .13)$. To assess children's perceived trustworthiness of the robot, we analyzed the corresponding subjective measures or items of trust. Again, we did not find any significant effect of regulatory focus on the dependent variables, *QAdv* $(U(N_{Prom} = 28, N_{Prev} = 26) = 363, z = -.02, p = .984, r = -.01)$, *QFolSug* $(U(N_{Prom} = 29, N_{Prev} = 26) = 379, z = .04, p = .969, r = .01)$, and *QSec* $(U(N_{Prom} = 28, N_{Prev} = 27) = 417, z = .68, p = .496, r = .09)$.

Other studies focused on assessing trust in social robots have suggested that children's perception of trust in a robot is rather inferred from initial impressions of competence and likability (Calvo-Barajas et al., 2020).

We ran Spearman's rank correlation analysis to examine if likability and competence were positively or negatively correlated with trust. The results are summarized in **Table 4**. The results revealed a positive significant correlation between the items of likability (QLik and QFri), competence (QIma and QHelp), and trust (QAvd and QFolSug). We found that the trust item QSec was significantly positively correlated with the items evaluated for perceiving competence (QIma and QHelp). This exploratory analysis shows that children's perception of the robot's likability and competence positively impacts participants' trust in the robot.

4.3 Children's Following of the Robot's Suggestions

RQ2 aimed to investigate the effect of regulatory focus design on children's acceptance of the robot's suggestions. To accomplish this, we defined five dependent variables, MCSug, MRSug,

[3] https://www.affectiva.com

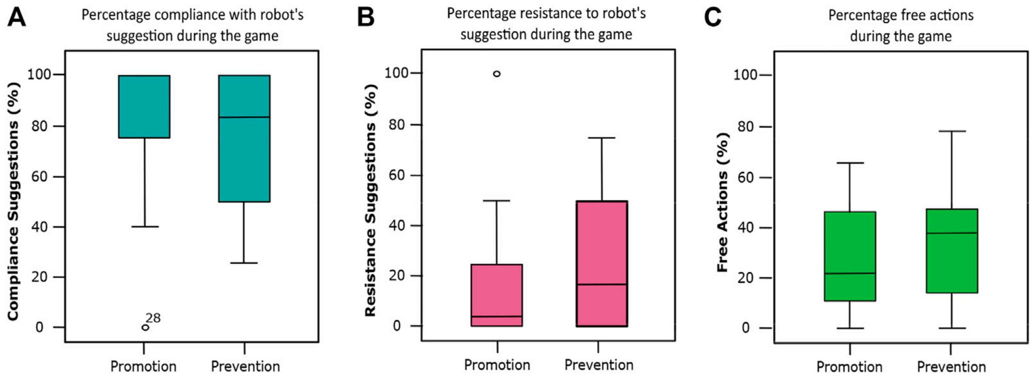


FIGURE 3 | (A) is the percentage of compliance suggestions, **(B)** is the resistance suggestions, and **(C)** is the free actions per condition during the interactive–collaborative game. There were no significant differences between conditions.

TABLE 5 | Spearman's rank correlations of subjective measures with objective measures.

	Subjective measures						
	Likability		Competency		Trust		
Objective measures	**Liking**	**Friendliness**	**Imagination**	**Helpfulness**	**Trust-advice**	**Follow-suggestions**	**Secrets**
Compliance with suggestions	–0.20	0.04	0.02	–0.39[a]	0.20	0.21	–0.09
Resistance to suggestions	0.21	–0.04	–0.01	0.43[a]	–0.19	–0.21	0.13
Compliance with requests	0.10	–0.02	–0.02	0.17	0.14	0.05	0.03
Resistance to requests	–0.10	0.02	0.02	–0.17	–0.14	–0.05	–0.03
Free actions	0.06	0.05	–0.13	0.23	–0.05	–0.06	–0.05

[a] $p < 0.05$.
[b] $p < 0.01$.

MCReq, MRReq, and MFAct, described in **Section 3.5.2**. To understand the effect of the condition on the dependent variables, the dependent variables were measured as frequencies rather than averages. In other words, we counted the number of times the child accepted the robot's suggestions. These measures were transformed into percentages for easier interpretation.

We ran a Kolmogorov–Smirnov test to check normality. All our dependent variables deviated significantly from normal. Thus, we ran a Mann–Whitney *U* test. The analysis did not reveal any significant difference between the two conditions for $MCSug\,(U\,(N_{Prom} = 28, N_{Prev} = 23) = 254, z = -1.34, p = .179, r = -.19)$, $MRSug\,(U\,(N_{Prom} = 28, N_{Prev} = 23) = 385, z = 1.26, p = .209, r = .17)$, $MCReq\,(U\,(N_{Prom} = 28, N_{Prev} = 24) = 355, z = .49, p = .627, r = .07)$, $MRReq\,(U\,(N_{Prom} = 28, N_{Prev} = 24) = 316, z = -.47, p = .627, r = -.07)$, $MFAct\,(U\,(N_{Prom} = 28, N_{Prev} = 24) = 388, z = .96, p = .339, r = .14)$. **Figure 3** shows the distribution of children's acceptance of and resistance to the robot's suggestions, as well as free actions. There was no significant difference between the conditions.

As an exploratory analysis, we investigated the relationship between subjective measures (i.e., QLik, QFri, QIma, QHelp, QAdv, QFolSug, and QSec) and objective measures (i.e., MCSug, MRSug, MCReq, MRReq, and MFAct). To do so, we ran Spearman's rank correlation analysis. The results are summarized in **Table 5**. We found that the perceived competence of the robot in terms of helpfulness was significantly negatively correlated with children's acceptance rate of the robot's suggestions. Conversely, children's perception of the robot's helpfulness significantly impacted children's resistance to following the robot's suggestions.

5 DISCUSSION

5.1 Effect of RFT on Perceived Likability, Competence, and Trustworthiness (RQ1)

We found that a regulatory focus design scenario affects children's perceptions of the likability of the robot. Our results suggest that children who interact with a social robot in a goal-oriented activity liked the robot more when it motivated them to achieve a goal to avoid risk (prevention-focused condition) than when it motivated them to get a reward (promotion-focused condition). This result is in line with previous work with virtual agents that found that prevention focus positively affects the likability of a virtual agent for users (Faur et al., 2015). One possible interpretation of these results is that the prevention-focused robot expressed verbal behaviors that communicated that it was scared of the explosion (**Section 3.3.1**) and, as a consequence, children might have associated these behaviors with a robot's vulnerability, leading to an increased perception of the likability of the robot. Prior work has found that vulnerable disclosures may drive more feelings of companionship with a robot in teenagers (Martelaro et al., 2016).

An interesting point of discussion concerns the relation between children's perception of the robot's likability measured post-interaction and their behavior during the priming game. While children rated the promotion robot as less likeable, behavioral data based on the facial expressions of emotion and engagement (see **Section 4.1**) showed that the promotion robot evoked more happiness and social engagement in children. However, this is not surprising as it is well established in social psychology that different types of measurement elicit different responses, and these different responses often do not correlate (Greenwald, 1990; De Houwer and Moors, 2007). It addresses an open question on the methods used to measure children's perceptions of and their social interaction with social robots. This is crucial when investigating child–robot relationship formation, as it has been shown that the type of measure (subjective vs. objective) influences how children interpret their social trust in and liking for a social robot (Stower et al., 2021).

Moreover, the results showed no significant difference in evaluations of perceived competence and trustworthiness for any of the items measured. This suggests that a regulatory focus design scenario does not directly affect the way children create a trust model of the robot. One possible explanation of this result could be that the robot's performance (equally for both conditions) had a stronger effect on children's perception of trust and competence with regard to the robot than the robot's expressiveness (i.e., happiness vs. fear), as responsiveness is associated with children's trust in a robot (van Straten et al., 2020). However, further investigation is needed to support this assertion.

Correlation analyses suggest that children's trust in a robot might be captured by impressions of likability and competence that the robot evokes. This result is in line with previous studies that suggest that these constructs are predictors of trust (Calvo-Barajas et al., 2020). In particular, we found that likability and competence positively affect the perception of competency-trust. This finding is surprising, because the literature has shown that relationship formation constructs (e.g., likability and friendliness) overlap with social-trust (Bernath and Feshbach, 1995; Stower et al., 2021; van Straten et al., 2020). In contrast, we found that the perceived competence the robot elicits has a positive effect on the children's consent to disclose their secrets to the robot. Again, this result is unexpected as measures of self-disclosure, keep-, and tell-secrets are associated with the definition of social-trust (i.e., "belief that the robot will keep its word or promises") (Stower et al., 2021).

We presume that a regulatory focus design scenario influences the way children build their trust model of a social robot, where the perceived likability and competence are positively significantly correlated with trust. One possible argument could be that in the prevention-focused condition, children experienced the need for security to reduce risk. Thus, when they accomplished the goal (i.e., getting out of the spaceship before it explodes), their perception of the helpfulness of the robot at avoiding a specific threat might have increased their social-trust in the robot. This preliminary explanation could be linked to the fact that prevention-focused behaviors are mediated by privacy concerns in adults (Wirtz and Lwin, 2009). Since our study is the first study of its nature, we need more evidence to support this claim.

Nevertheless, the conceptualization and operationalization of trust are challenging, especially in cHRI, as its definition differs among individuals. Hence, we considered the multifaceted property of trust as a key element to be exploited for a better understanding of how children make judgments of trustworthiness. The design of tailored methods and measures to capture children's trust in robots is gaining the attention of researchers from different fields to reduce the heterogeneity of this construct among studies (van Straten et al., 2020). We hope that our findings provide insights that can be used to build on the conceptualization of children's trust and its implications with regard to the relationship with a robot.

5.2 Effect of RFT on Children's Following of the Robot's Suggestions (RQ2)

Concerning children's willingness to accept or resist the robot's suggestions, we found that a regulatory focus design scenario does not significantly affect children's rates of acceptance of the robot's suggestions and requests. Even though we did not find any significant difference, we noticed that on average the prevention-focused condition elicits higher resistance in children to following the robot's suggestions, whereas the promotion-focused condition seems to influence a higher reliance on the robot's suggestions, which is associated with higher competency trust in the robot. This preliminary result aligns with prior research in psychology, as it suggests that promotion-focused individuals are open to new ideas and experiences (Friedman and Förster, 2001). However, our results were not statistically significant, and more investigation is needed to claim this statement. Nevertheless, we believe that our findings could be beneficial for further studies as they provide new insights into the design of the robot's affective behaviors based on RFT to elicit positive emotions, a paradigm that has not been studied before in cHRI. Therefore, it could be beneficial for child–robot relationship formation, especially in the domain of child–robot educational interactions.

Moreover, we find the ceiling effect observed in children's compliance with the robot's suggestions interesting but not surprising. In several cases, children accepted all the suggestions, taking into account that some of them were wrong. These results raise opportunities, but also concerns, regarding the use of social robots as learning companions for young children, as has been presented in prior work (Vollmer et al., 2018).

Finally, the exploratory analysis revealed that the perceived helpfulness of the robot negatively impacted the children's compliance with the robot's suggestions. This result is confounding, as we would have expected that children who perceived the robot to be more helpful would be more likely to follow the robot's suggestions. However, this result should be interpreted with caution, as helpfulness was assessed as a subjective post-interaction measure, whereas children's compliance/resistance with/to the robot's suggestions was assessed as an objective measure during the interaction. On one hand, we presume that other parts of the activity could have influenced children's judgment of the robot's helpfulness. On the other hand, previous research studies have found that subjective and objective measures elicited different responses when evaluating children's competency trust in a social robot (Stower et al., 2021).

Overall, our results suggested that objective measures are not always positively correlated with subjective measures. However, as indicated in the study by Belpaeme et al. (2013), it is crucial to validate if the desired outcome is captured by the proposed objective measure, as some constructs are harder to measure than others. To explore the relationship between subjective and objective measures, we would like to further investigate whether "following the robot's suggestions" is an appropriate and reliable measure to capture children's competency trust in the robot in a regulatory focus priming scenario without having the storytelling activity component, as proposed in our pilot study Calvo-Barajas et al. (2021).

5.3 Limitations and Future Work

One of the limitations of this study is that we were not able to explore the effects of a regulatory focus design scenario on children's compliance with right and wrong suggestions. The nature of the interaction did not allow us to have the same amount of right and wrong suggestions between conditions. However, the exploration of these effects would be an interesting topic for future research. As such, we aim at increasing the number of times children have to comply with or resist the robot's suggestion, and this might also improve the inter-rater agreement score.

Another limitation is that we could not fully explore whether the manipulation of the regulatory focus induced negative feelings of stress in the participants. We were only able to measure differences in terms of expressions of happiness and social engagement between the two conditions. In future work, the measurement of electrodermal activity (EDA) could be considered to analyze children's stress level.

In this study, we were interested in investigating regulatory focus theory as a priming strategy rather than exploring matching as the regulatory fit theory suggests. Therefore, we did not assess children's regulatory orientation. Nevertheless, we believe that this is an interesting topic for further investigation, as individual differences might influence children's social interactions and relationship with robots (Shahid et al., 2014). In addition, it would be worthwhile to explore different methods of the robot's adaptation in cHRI.

As we discussed before, we did not find any significant difference in regulatory focus design on child–robot relationship formation. To provide more insights into the implementation of RFT as a technique to be used in cHRI in an educational context, it would be interesting to study whether the robot's presence influences the way children interact in a goal-oriented task based on RFT by introducing a control condition.

Finally, a parallel line of investigation, but outside the scope of this article, includes the exploration of whether and how RFT affects creativity performance in interactive storytelling. To do so, we aim to evaluate narrative creativity measures in children's stories before and after the priming activity.

6 CONCLUSION

In this article, we presented a novel user study investigating the effects of a regulatory focus design scenario on children's perception of likability, competence, and trustworthiness of a social robot. Besides, we evaluated the effect of a regulatory focus design scenario on children's compliance with the robot's suggestions.

We found that a regulatory focus design scenario significantly affected children's perception of the likability of the robot, while perceived competence and trustworthiness did not change between conditions. Similarly, the motivation to achieve a goal did not significantly affect the way children followed the robot's suggestions. Nevertheless, on average, the prevention-focused robot increased children's resistance to following suggestions. Interestingly, the items used to capture children's trust in a robot are correlated among them, suggesting that trust may be inferred by constructs of social cognition and social learning.

These findings are relevant to the study of trust in cHRI, as they provide new evidence on the effect of strategies based on RFT on perceived trust in a robot in an educational scenario, and they highlight the relevance of the multidimensional nature of trust when evaluating children's judgments of trustworthiness of a social robot.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comissão de Ética do Instituto Superior Técnico Lisboa (CE-IST), Lisbon, Portugal. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

NC-B wrote the initial draft of this article, which included data analysis, the creation of an initial article structure, the creation of the coding scheme for objective measures, and leading the research discussions on iterative article improvements.

Together with ME, she designed and conducted the user study and carried out most of the annotations. NC-B, GP, and GC have initiated discussions on children's trust in social robots. All the authors provided feedback on the article's structure and wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been funded by the European Commission Horizon 2020 Research and Innovation Program under Grant Agreement no. 765955. The authors are solely responsible for the content of this publication.

REFERENCES

Agrigoroaie, R., Ciocirlan, S.-D., and Tapus, A. (2020). In the Wild Hri Scenario: Influence of Regulatory Focus Theory. *Front. Robot. AI* 7, 58. doi:10.3389/frobt.2020.00058

Baas, M., De Dreu, C. K. W., and Nijstad, B. A. (2008). A Meta-Analysis of 25 Years of Mood-Creativity Research: Hedonic Tone, Activation, or Regulatory Focus?. *Psychol. Bull.* 134, 779–806. doi:10.1037/a0012815

Belpaeme, T., Baxter, P., de Greeff, J., Kennedy, J., Read, R., Looije, R., et al. (2013). "Child-robot Interaction: Perspectives and Challenges," in *Social Robotics*. Editors G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, and U. Leonards (Cham: Springer International Publishing), 452–459. doi:10.1007/978-3-319-02675-6_45

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social Robots for Education: A Review. *Sci. Robot.* 3, eaat5954. doi:10.1126/scirobotics.aat5954

Bernath, M. S., and Feshbach, N. D. (1995). Children's Trust: Theory, Assessment, Development, and Research Directions. *Appl. Prev. Psychol.* 4, 1–19. doi:10.1016/S0962-1849(05)80048-4

Bethel, C. L., Henkel, Z., Stives, K., May, D. C., Eakin, D. K., Pilkinton, M., et al. (2016). "Using Robots to Interview Children about Bullying: Lessons Learned from an Exploratory Study," in 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN); New York, NY, USA, 26-31 Aug. 2016 (IEEE Press), 712–717

Beuk, F., and Basadur, T. (2016). Regulatory Focus, Task Engagement and Divergent Thinking. *Creativity Innovation Manage.* 25, 199–210. doi:10.1111/caim.12182

Breazeal, C., Harris, P. L., DeSteno, D., Kory Westlund, J. M., Dickens, L., and Jeong, S. (2016). Young Children Treat Robots as Informants. *Top. Cogn. Sci.* 8, 481–491. doi:10.1111/tops.12192

Calvo, N., Elgarf, M., Perugia, G., Peters, C., and Castellano, G. (2020). "Can a Social Robot Be Persuasive without Losing Children's Trust?," in HRI'20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (New York, NY, USA: Association for Computing Machinery), 157–159. doi:10.1145/3371382.3378272

Calvo-Barajas, N., Perugia, G., and Castellano, G. (2020). "The Effects of Robot's Facial Expressions on Children's First Impressions of Trustworthiness," in 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 165–171. doi:10.1109/RO-MAN47096.2020.9223456

Calvo-Barajas, N., Perugia, G., and Castellano, G. (2021). "The Effects of Motivational Strategies and Goal Attainment on Children's Trust in a Virtual Social Robot: A Pilot Study," in Interaction Design and Children (IDC'21). New York, NY, USA: Association for Computing Machinery, 537–541.

Cameron, D., de Saille, S., Collins, E. C., Aitken, J. M., Cheung, H., Chua, A., et al. (2020). The Effect of Social-Cognitive Recovery Strategies on Likability, Capability and Trust in Social Robots. *Comput. Hum. Behav.* 114, 106561. doi:10.1016/j.chb.2020.106561

Carpenter, M., and Nielsen, M. (2008). Tools, Tv, and Trust: Introduction to the Special Issue on Imitation in Typically-Developing Children. *J. Exp. Child Psychol.* 101 (4), 225–227. doi:10.1016/j.jecp.2008.09.005

Chen, H., Park, H. W., Zhang, X., and Breazeal, C. (2020). "Impact of Interaction Context on the Student Affect-Learning Relationship in Child-Robot Interaction," in Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction HRI'20 (New York, NY, USA: Association for Computing Machinery), 389–397. doi:10.1145/3319502.3374822

Cornwell, J. F. M., and Higgins, E. T. (2016). Eager Feelings and Vigilant Reasons: Regulatory Focus Differences in Judging Moral Wrongs. *J. Exp. Psychol. Gen.* 145, 338–355. doi:10.1037/xge0000136

Cruz-Maya, A., Agrigoroaie, R., and Tapus, A. (2017). "Improving User's Performance by Motivation: Matching Robot Interaction Strategy with User's Regulatory State," in International Conference on Social Robotics. Editors A. Kheddar, et al. (Cham: Springer), 10652, 464–473. doi:10.1007/978-3-319-70022-9_46

Cruz-Maya, A., and Tapus, A. (2018). "Adapting Robot Behavior Using Regulatory Focus Theory, User Physiological State and Task-Performance Information," in 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (IEEE), 644–651

Dawe, J., Sutherland, C., Barco, A., and Broadbent, E. (2019). Can Social Robots Help Children in Healthcare Contexts? a Scoping Review. *Bmjpo* 3, e000371. doi:10.1136/bmjpo-2018-000371

De Houwer, J., and Moors, A. (2007). "How to Define and Examine the Implicitness of Implicit Measures," in *Implicit Measures of Attitudes: Procedures and Controversies*. Editors B. Wittenbrink and N. Schwarz The Guilford Press, 179–194

Elgarf, M., Calvo-Barajas, N., Paiva, A., Castellano, G., and Peters, C. (2021). "Reward Seeking or Loss Aversion? Impact of Regulatory Focus Theory on Emotional Induction in Children and Their Behavior Towards a Social Robot," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (New York, NY, USA: Association for Computing Machinery), 1–11. doi:10.1145/3411764.3445486

Faur, C., Martin, J.-C., and Clavel, C. (2015). "Matching Artificial Agents'and Users' Personalities: Designing Agents with Regulatory-Focus and Testing the Regulatory Fit Effect," in Annual Meeting of the Cognitive Science Society (CogSci). Pasadena, California, United States

Friedman, R. S., and Förster, J. (2001). The Effects of Promotion and Prevention Cues on Creativity. *J. Personal. Soc. Psychol.* 81, 1001–1013. doi:10.1037/0022-3514.81.6.1001

Garcia-Garcia, J. M., Penichet, V. M., and Lozano, M. D. (2017). "Emotion Detection: a Technology Review," in Proceedings of the XVIII international conference on human computer interaction. New York, NY, 1–8. Interacción '17. doi:10.1145/3123818.3123852

Geiskkovitch, D. Y., Thiessen, R., Young, J. E., and Glenwright, M. R. (2019). "What? That's Not a Chair!: How Robot Informational Errors Affect Children's Trust towards Robots," in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE Press), 48–56

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., et al. (2016). "Affective Personalization of a Social Robot Tutor for Children's Second Language Skills," in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI Press), 3951–3957

Greenwald, A. G. (1990). What Cognitive Representations Underlie Social Attitudes? *Bull. Psychon. Soc.* 28, 254–260. doi:10.3758/bf03334018

Groom, V., Srinivasan, V., Bethel, C. L., Murphy, R., Dole, L., and Nass, C. (2011). "Responses to Robot Social Roles and Social Role Framing," in 2011 International Conference on Collaboration Technologies and Systems (CTS) (IEEE), 194–203

Heerink, M., Kröse, B., Evers, V., and Wielinga, B. (2010). Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model. *Int. J. Soc. Rob.* 2, 361–375. doi:10.1007/s12369-010-0068-5

Higgins, E. T. (1997). Beyond Pleasure and Pain. *Am. Psychol.* 52, 1280–1300. doi:10.1037/0003-066x.52.12.1280

Higgins, E. T. (1998). Promotion and Prevention: Regulatory Focus as a Motivational Principle. *Adv. Exp. Soc. Psychol.* 30, 1–46. doi:10.1016/s0065-2601(08)60381-0

Higgins, E. T., and Cornwell, J. F. M. (2016). Securing Foundations and Advancing Frontiers: Prevention and Promotion Effects on Judgment & Decision Making. *Organizational Behav. Hum. Decis. Process.* 136, 56–67. doi:10.1016/j.obhdp.2016.04.005

Higgins, E. T. (2005). Value from Regulatory Fit. *Curr. Dir. Psychol. Sci.* 14, 209–213. doi:10.1111/j.0963-7214.2005.00366.x

Howes, C., and Ritchie, S. (2002). *A Matter of Trust: Connecting Teachers and Learners in the Early Childhood Classroom*. New York, NY: Teachers College Press, Vol. 84

Johnston, A. M., Mills, C. M., and Landrum, A. R. (2015). How Do Children Weigh Competence and Benevolence when Deciding Whom to Trust? *Cognition* 144, 76–90. doi:10.1016/j.cognition.2015.07.015

Kahn, P. H., Jr, Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., et al. (2012). "Robovie, You'll Have to Go into the Closet Now": Children's Social and Moral Relationships with a Humanoid Robot. *Dev. Psychol.* 48, 303–314. doi:10.1037/a0027033

Kedzierski, J., Muszyński, R., Zoll, C., Oleksy, A., and Frontkiewicz, M. (2013). Emys—emotive Head of a Social Robot. *Int. J. Soc. Robotics* 5, 237–249. doi:10.1007/s12369-013-0183-1

Keller, J., Mayo, R., Greifeneder, R., and Pfattheicher, S. (2015). Regulatory Focus and Generalized Trust: The Impact of Prevention-Focused Self-Regulation on Trusting Others. *Front. Psychol.* 6, 254. doi:10.3389/fpsyg.2015.00254

Kennedy, J., Baxter, P., and Belpaeme, T. (2015). "The Robot Who Tried Too Hard," in HRI'15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (New York, NY, USA: Association for Computing Machinery), 67–74. doi:10.1145/2696454.2696457

Kim, P. H., Dirks, K. T., and Cooper, C. D. (2009). The Repair of Trust: A Dynamic Bilateral Perspective and Multilevel Conceptualization. *Amr* 34, 401–422. doi:10.5465/amr.2009.40631887

Kok, B. C., and Soh, H. (2020). Trust in Robots: Challenges and Opportunities. *Curr. Robot Rep.* 1, 297–309. doi:10.1007/s43154-020-00029-y

Kory Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., et al. (2017). Flat vs. Expressive Storytelling: Young Children's Learning and Retention of a Social Robot's Narrative. *Front. Hum. Neurosci.* 11, 1–20. doi:10.3389/fnhum.2017.00295

Landis, J. R., and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *biometrics* 33, 159–174. doi:10.2307/2529310

Landrum, A. R., Mills, C. M., and Johnston, A. M. (2013). When Do Children Trust the Expert? Benevolence Information Influences Children's Trust More Than Expertise. *Dev. Sci.* 16, 622–638. doi:10.1111/desc.12059

Lee, J. D., and See, K. A. (2004). Trust in Automation: Designing for Appropriate reliance. *hfes* 46, 50–80. doi:10.1518/hfes.46.1.50.30392

Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. (2014). Empathic Robots for Long-Term Interaction. *Int. J. Soc. Robotics* 6, 329–341. doi:10.1007/s12369-014-0227-1

Leite, I., McCoy, M., Lohani, M., Ullman, D., Salomons, N., Stokes, C., Rivers, S., and Scassellati, B. (2015). "Emotional Storytelling in the Classroom," in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (New York, NY: Association for Computing Machinery), 75–82. doi:10.1145/2696454.2696481

Liberman, N., Molden, D. C., Idson, L. C., and Higgins, E. T. (2001). Promotion and Prevention Focus on Alternative Hypotheses: Implications for Attributional Functions. *J. Personal. Soc. Psychol.* 80, 5–18. doi:10.1037/0022-3514.80.1.5

Madsen, M., and Gregor, S. (2000). "Measuring Human-Computer Trust," in 11th australasian conference on information systems (Citeseer), 53, 6–8.

Malle, B. F., and Ullman, D. (2021). "Chapter 1 - A Multidimensional conception and Measure of Human-Robot Trust," in *Trust in Human-Robot Interaction*. Editors C. S. Nam and J. B. Lyons (Academic Press), 3–25. doi:10.1016/B978-0-12-819472-0.00001-0

Martelaro, N., Nneji, V. C., Ju, W., and Hinds, P. (2016). "Tell Me More Designing Hri to Encourage More Trust, Disclosure, and Companionship," in The 11th ACM/IEEE International Conference on Human-Robot Interaction HRI 16 (IEEE Press), 181–188

McAllister, D. J. (1995). Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Acad. Manag. J.* 38, 24–59. doi:10.2307/256727

Neumann, R. (2000). The Causal Influences of Attributions on Emotions: A Procedural Priming Approach. *Psychol. Sci.* 11, 179–182. doi:10.1111/1467-9280.00238

Perugia, G., Díaz-Boladeras, M., Català-Mallofré, A., Barakova, E. I., and Rauterberg, M. (2020). Engage-dem: A Model of Engagement of People with Dementia. *IEEE Trans. Affective Comput.*, 1. doi:10.1109/taffc.2020.2980275

Ryan, R. M., and Grolnick, W. S. (1986). Origins and Pawns in the Classroom: Self-Report and Projective Assessments of Individual Differences in Children's Perceptions. *J. Personal. Soc. Psychol.* 50, 550–558. doi:10.1037/0022-3514.50.3.550

Saerbeck, M., Schut, T., Bartneck, C., and Janse, M. D. (2010). "Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor," in Proceedings of the 28th international conference on Human factors in computing systems - CHI'10 (New York, NY, USA: Association for Computing Machinery), 1613–1622

Shahid, S., Krahmer, E., and Swerts, M. (2014). Child-robot Interaction across Cultures: How Does Playing a Game with a Social Robot Compare to Playing a Game Alone or with a Friend? *Comput. Hum. Behav.* 40, 86–100. doi:10.1016/j.chb.2014.07.043

Stower, R., Calvo-Barajas, N., Castellano, G., and Kappas, A. (2021). A Meta-Analysis on Children's Trust in Social Robots. *Int. J. Soc. Robotics*. doi:10.1007/s12369-020-00736-8

Straten, C. L. v., Kühne, R., Peter, J., de Jong, C., and Barco, A. (2020). Closeness, Trust, and Perceived Social Support in Child-Robot Relationship Formation. *Is* 21, 57–84. doi:10.1075/is.18052.str

Tielman, M., Neerincx, M., Meyer, J.-J., and Looije, R. (2014). "Adaptive Emotional Expression in Robot-Child Interaction," in HRI'14: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (New York, NY, USA: Association for Computing Machinery), 407–414. doi:10.1145/2559636.2559663

van Straten, C. L., Peter, J., and Kühne, R. (2020). Child-Robot Relationship Formation: A Narrative Review of Empirical Research. *Int. J. Soc. Robotics* 12, 325–344. doi:10.1007/s12369-019-00569-0

van Straten, C. L., Peter, J., Kühne, R., de Jong, C., and Barco, A. (2018). "Technological and Interpersonal Trust in Child-Robot Interaction: An Exploratory Study," in HAI 18 Proceedings of the 6th International Conference on Human-Agent Interaction (New York, NY, USA: Association for Computing Machinery), 253–259

Verhagen, J., Berghe, R. v. d., Oudgenoeg-Paz, O., Küntay, A., and Leseman, P. (2019). Children's reliance on the Non-verbal Cues of a Robot versus a Human. *PLoS One* 14, e0217833. doi:10.1371/journal.pone.0217833

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., et al. (2019). "Second Language Tutoring Using Social Robots: A Large-Scale Study," in Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) IEEE Press, 497–505

Vollmer, A.-L., Read, R., Trippas, D., and Belpaeme, T. (2018). Children Conform, Adults Resist: A Robot Group Induced Peer Pressure on Normative Social Conformity. *Sci. Robot.* 3, eaat7111. doi:10.1126/scirobotics.aat7111

Wirtz, J., and Lwin, M. O. (2009). Regulatory Focus Theory, Trust, and Privacy Concern. *J. Serv. Res.* 12, 190–207. doi:10.1177/1094670509335772

Witt, P. L., Wheeless, L. R., and Allen, M. (2004). A Meta-analytical Review of the Relationship between Teacher Immediacy and Student Learning. *Commun. Monogr.* 71, 184–207. doi:10.1080/036452042000228054

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership