International Journal of Social Robotics (2021) 13:1013–1031
https://doi.org/10.1007/s12369-020-00688-z

Is the Social Desirability Effect in Human–Robot Interaction overestimated? A Conceptual Replication Study Indicates Less Robust Effects

Benedikt Leichtmann[1] · **Verena Nitsch**[1]

Accepted: 29 July 2020 / Published online: 7 September 2020

Abstract

The "Computers are social actors" (CASA) assumption (Nass and Moon in J Soc Issues 56:81–103, 2000. https://doi.org/10.1111/0022-4537.00153) states that humans apply social norms and expectations to technical devices. One such norm is to distort one's own response in a socially desirable direction during interviews. However, findings for such an effect are mixed in the literature. Therefore, a new study on the effect of social desirability bias in human–robot evaluation was conducted, aiming for a conceptual replication of previous findings. In a between-subject laboratory experiment, $N = 107$ participants had to evaluate the robot and the interaction quality after a short conversation in three different groups: In one group, the evaluation was conducted using (1) the same robot of the former interaction, (2) a different robot, (3) a tablet computer. According to the CASA assumption, it was expected, that evaluations on likability and quality of interaction, are higher in the condition with the same robot conducting the evaluation, compared to a different robot or a tablet computer because robots are treated as social actors and hence humans distort ratings in a socially desirable direction. Based on previous findings, we expected robots to evoke higher anthropomorphism and feelings of social presence compared to the tablet computer as potential explanation. However, the data did not support the hypotheses. Low sample size, low statistical power, lack of measurement validation and other problems that could lead to an overestimation of effect sizes—in this study and the literature in general—are discussed in light of the replicability crisis.

Keywords Social desirability bias · Social robotics · Anthropomorphism · Replicability crisis · Media equation · Social presence

1 Introduction

For product improvement and development, user evaluation is key. Certain technological products such as video games or robots, can directly inquire about the user's subjective use experience without the involvement of a third party, such as a human interviewer. If human service is evaluated by humans (for example, a restaurant's chef asking how the food is), psychological research shows that the answer is on average distorted in a socially desirable direction [43,49] (Most people would tell the chef that they like the food, even if it was not that good.). This distortion of answers is seen as a rule of politeness, a social norm, that is "jointly negotiated rules of social behavior" [11, p. 151].

Social desirability bias is an intensively studied psychological phenomenon and "refers to the tendency by respondents, under some conditions and modes of administration, to answer questions in a more socially desirable direction than they would under other conditions or modes of administration" [55, p. 755]. The level of this distortion (response bias) can be obtained by the mean difference in scores on socially sensitive questions. These include questions in which the answer is distorted toward a prevailing social norm or value (normative response bias), or questions, in which the answer is distorted toward what is expected to be preferred by the interviewer and thus following social rules

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12369-020-00688-z) contains supplementary material, which is available to authorized users.

✉ Benedikt Leichtmann
b.leichtmann@iaw.rwth-aachen.de

Verena Nitsch
v.nitsch@iaw.rwth-aachen.de

[1] Institute of Industrial Engineering and Ergonomics, RWTH Aachen University, Aachen, Germany

of politeness (interviewer-based bias) [48]. The latter would mean to rate a person as attractive or likable independently of ones actual opinion because offending other people by evaluative statements is perceived as a socially unacceptable form of interaction. Thus, avoiding this offense can be seen as a maxim of politeness. One possible solution to avoid these response biases in interviews is to reduce the "social nature" of the interview setting. Researchers have tried to obtain this goal by computerized interviews compared to traditional ones. Psychological research shows, using a computer might make the human feel more anonymous and reduce social pressure [55].

However, scholars have argued, that this social desirability bias might also occur in human–technology interaction scenarios to a certain extent [see 47,48]. In certain contexts, human users apply social norms such as response biases (likely unintentionally) and thus treat technology as social actors. This assumption leads to various implications for human–robot interaction designs: For instance, robots that conduct their own user evaluations, would result in distorted response patterns. This is especially conceivable for social robots because they display social characteristics and are implemented in social settings leading to a higher "social nature" than other devices made for other purposes. Following these assumptions, the evaluation of social robots should rather be examined by other devices, such as a separate website that evokes less social attributions. However, context factors, the characteristics of the evaluator, as well as the object of evaluation, and consequently underlying psychological mechanisms that describe how certain factors are connected and lead to certain observations is still under research [15,68].

The aim of this article is thus two-fold: first, this article aims to contribute toward questions on the robustness of social desirability effects by adding further insights from a conceptual replication of a study by Nass et al. [48] and other replications in different contexts. Social-psychology has been facing a serious replicability crisis that also affects social robotics and thus more tests on the robustness of already established effects need to be performed. The second aim is to expand the body of research by applying this hypothesis to human–robot interaction. While this effect of social desirability has been tested with different devices such as computers [48] or smartphones [9], we are not aware of a study testing this with robots. Instead of using computers, this study tested the social desirability effect on the evaluation of human–robot interaction. In doing so we test differences in anthropomorphism and perceived social presence as potential key factors.

2 Theoretical Background and Related Work

In order to understand the motivation to conduct another study on effects that have already been studied in a similar way before, an overview of a recent replicability crisis is needed, that especially affected social psychology in recent years calling into question many effects that were thought to be well-established and raising doubt on their robustness. After a more general discussion, first examples are given that this is a potential problem in human–robot interaction studies, too. Second, a brief literature overview is given on the social desirability effect in human–technology interaction to show why one would even assume to find such an effect in human–robot interaction. Therefore, two connected theories are discussed to explain this effect. The "computers are social actors" (CASA) hypothesis, or the "media equation" hypothesis [47] describe the basic idea that technology is treated similar to social beings, while the body of research on anthropomorphism [15] delivers more fine grained hypotheses on when to expect larger social desirability effects. Based on these theories, we go into more detail on what is actually known upto now on the social-desirability effect concerning technology in a third step—what are weaknesses and what can be learned for such a social-desirability effect in human–robot interaction, leading to our research question and hypotheses.

2.1 A Century of Crisises: Problems of Replicability, Validity and Theory

Before going into details of the actual phenomenon of social desirability in human–technology interaction, one might wonder why it is of great importance to conduct another experiment on a phenomenon that had been covered in the past several times such as the social desirability effect in human–computer interaction [48] or other technology such as smartphones [9]. In the last couple of years, well-established psychological phenomena underwent several large-scale attempts to replicate classical psychological laboratory experiments by different research groups [see 36,50]. However, results showed a substantial decline in effect sizes compared to the original findings—the findings could not be replicated. As a consequence, disciplines such as neuroscience [30] or social psychology [63] faced a huge crisis, also known as the replicability crisis. Problems were manifold.

One of the main reasons for this problem are underpowered studies [42]. Statistical power is the probability of correctly rejecting the null hypothesis given that the alternative hypothesis is true. Power is a function of effect size, sample size and alpha-level [38]. Therefore, if the (true) effect sizes are small in general, as it is rather the rule for social-psychological phenomena [58], or the variance of true effects is large, the average sample sizes are too small and thus studies have low power. Low powered studies in turn have a low probability of finding these true effects, producing more false positive results [8]. Even worse, effect sizes in under-powered studies are often overestimated, a problem also known as the "winner's curse" [8].

Hence, studies with low power can lead to false conclusions on what is actually "known". These false conclusions are additionally distorted by so called questionable research practices (QPRs) that increase the Type-I error rate dramatically and thus produce more false-positive results. These practices include, but are not limited to multiple testing of hypotheses without taking into account the family-wise error rate, hypothesizing after the results are known (HARKing), or controlling for variables a posteriori without theoretical legitimization[1] [60,71]. Low statistical power and questionable research practices can lead to false conclusions. Further, publication bias (publishing statistically significant results only or results that are in favor of a certain hypothesis, leaving results with no effect or non-expected findings unnoticed) distorts the results of meta-analyses and reviews leading to overestimation of effect sizes in the literature.

Besides problems of statistical data analysis, sample size justification and publication habits, at the core of the replicability crisis are problems of validation and lack of theory. Before statistical analysis, measurement comes first. Psychological studies mostly deal with variables that cannot be observed directly, but are latent [for a detailed discussion see 5]. Measurement methods only reflect a construct in a probabilistic way and thus methods vary in the degree of accuracy. That means validity needs to be shown by a validation process [5,12]. However, Hussey and Hughes [29] show that measures in psychology do not always undergo a thorough validation process, measures might thus be not valid or only of poor accuracy (that is, a measure reflects a big amount of noise compared to the construct in question). However, even if a measurement method had been shown to be valid in one situation, that does not mean that it is valid in others. For example, while a method might have been shown to be valid in the interaction with a mechanic-looking industrial robot, it might not work so well in interaction-situations with humanoid robots. Or while some methods might work today—with many people interacting with a robot for the first time—it may happen to be a less suitable measure in 50 years with more and more people experiencing situations of human–robot interaction.

The basis for validity is a good theoretical grounding. For a measure to be valid the variation in indices must be caused by variations in the theoretical construct, and this theoretical construct must exist in the first place. So there is theory needed to explain "what happens between the attribute and the test scores" [6, p. 1067]. Further, these constructs need to be integrated into a nomological network [12], a system of laws relating constructs to other constructs, and in theoretical frameworks as a body of connected theories [46], that can help to narrow down the space of numerous potential influential factors systematically. However, many studies in psychology seem to be based on guesswork or folk intuitions [46].

In sum, mixed findings on a certain topic can thus be due to (1) the absence of a real effect and thus due to random noise, (2) true differences of effects based on mediating or moderating factors yet not taken into account, (3) invalid or insufficiently valid measurements, or (4) varying methodological quality concerning research design, sample size, and statistical analysis, as well as publication bias.

These problems have not gotten much attention in the HRI community, yet. However, the discussion on replicability problems has been growing in the human–robot interaction (HRI) community in recent years [see for example 3,31]. For example, Irfan et al. [31] reported two failed attempts to replicate social facilitation effects from social psychology in HRI settings and discussed their findings in the context of the replicability crisis. A recent review and meta-analysis by Leichtmann and Nitsch [39] revealed similar problems of low sample size, low statistical power and a lack of theory in proxemics studies in HRI settings.

These are just first examples of disucssions of such problems of replicability, validity and theory within the HRI domain. Effects in HRI contexts need to be tested for robustness more carefully. As shown in this section solution to this problem is a stronger theoretical stance, justification of sample size based on power-analysis, more careful validation of measurement tools, the avoidance of questionable research practices and stronger transparency in future studies.

2.2 Robots as Social Actors and Anthropomorphism

As mentioned in the introduction of this article, the goals are to test the robustness of the social desirability effect in HRI in conducting a conceptual replication of the effect and to extend the findings to human–robot interaction. However, the social desirability effect stems from social psychology and thus is not a genuinely technological phenomenon. In order to understand why one would expect to find such an effect in certain situations of human–technology interaction, two theoretical approaches need to be introduced. More than 20 years ago, Nass and colleagues (for an overview see [47]) conducted several experiments to illustrate their claim that

[1] For example, let's say a hypothesis could not be accepted based on a statistical test, but after controlling for several variables, one finds out that somehow in the data at hand the hypothesis only holds for left-handed men older than 50 for no reason. Accepting this purely data-driven result as a "finding" would follow a degenerating research line in the Lakatosian sense [37].

social and natural rules from interpersonal interaction apply *equally well* to media, an assumption named the media equation [51]. In a row of experiments (i.e. $k = 35$ studies were mentioned in [51], many more built upon this idea in the literature), they thus tested the hypothesis that human users treat technology such as computers or robots as social actors that seemed to support this claim. According to this CASA hypothesis ("computers are social actors"), people subconsciously apply social rules and expectations to computers (or more generally to technology). This effect is hypothesized to be moderated by contextual cues that trigger certain scripts and labels leading to social behavior. Based on this assumption, researchers have conducted various experiments, in which findings of social psychology about human–human interaction were used to predict behavior also in human–robot interaction with the expectation to find similar results [17,19,45,48,73]. One of several examples of the media equation assumption was the the social desirability bias that was claimed to equally occur in human–computer interaction as in human–human social interaction [47,48].

However, the attempt to test this assumption turned out not always to be successful [28,31]. Irfan et al. [31] were not able to replicate social psychological effects in human–robot interaction that would have been expected if robots were treated similarly as human beings. Another example is personal space in HRI, in which comfort distance toward a robot would be assumed to be similar as toward other humans under the CASA assumption. However, a recent review and meta-analysis on personal space in HRI, for example, indicates a more mixed picture [39].

The reader should note that based on these findings, the CASA assmuption is not proven, but it is not being refuted either. This just indicates that the CASA assumption might just be an over-generalization: technology might not be treated as a social being by *every* person in *every* situation. Thus, it needs to be moved beyond the question if CASA-based effects exist or not, but rather how strong these effects are depending on situational characteristics and thus how practically important such an effect is. Possible reasons for a mixed picture might be methodological problems such as small sample sizes (see Sect. 2.1), but also that the CASA theory does not hold for every situation or that effects are not equally strong.

Thus more detailed theories and subsequent key factors need to be identified that can explain such a variance in effects. Researchers have varied certain context characteristics in order to identify factors, which cause variation in applying social norms in human–technology interaction situations. One prominent construct in this context is anthropomorphism. Although there are differences in definition [14,57] and the construct seems often vaguely defined, anthropomorphism in general means "[…] imbuing the imagined or real behavior of nonhuman agents with humanlike characteristics, motivations, intentions, and emotions" [15, pp. 864–865]. This includes all nonhuman agents such as animals, nature, technology or virtual agents. This tendency is a continuum and varies as a function of dispositional, situational, developmental, and cultural influences [15]. Thus, people differ in the degree they anthropomorphize [69], and some situations evoke more anthropomorphism than others [18]. Oftentimes anthropomorphism is portrayed as a conscious process. Thus, Reeves and Nass [51] excluded anthropomorphism as a potential explanation for CASA-related effects by arguing that these CASA-related effects are sub-conscious. However, there is no need to understand anthropomorphism as a conscious process only [74]. Zlotowski et al. [74] explain this divergent results of low explicitly reported anthropomorphism on one side and actual behavior in favor of the CASA theory on the other side using dual-process models. Such dual process models rooted in social cognition research differentiate between an implicit process that is described as fast, unconscious and requiring only little effort resulting in automatic responses, and an explicit process described as slow, conscious and requiring higher amount of effort resulting in more controlled responses [74]. Similarly, other general social-cognitive models can explain such divergent observation using dual-processes [see for example 22,23] or a more integrated system, a "single evaluative system in which there are many functionally discrete computational processes" [13, p. 752] such as the iterative reprocessing model by Cunningham et al. [13].

While these models are process-oriented attempts, other researchers have focused more on the identification of factors that would lead to more or less anthropomorphism. Such factors include for example more motivational top-down factors, such as the three underlying psychological determinants proposed by Epley et al. [15]: According to this, people anthropomorphize (1) when anthropomorphic knowledge is accessible and applicable, (2) as a strategy to understand and predict the behavior of nonhuman agents (effectance motivation), (3) when the desire for social contact is high (sociality motivation). Research on these underlying mechanisms is still in its infancy, but first attempts already seem to support the theory so far [16,18].

Especially in human–technology interaction, most researchers, however, have focused on bottom-up design-oriented features of the robot, that would activate representations of human-like features and subsequently lead to an activation of other related human-like characteristics such as mind-attribution. Such features serving as social cues [1] include voice [19,70] or movement [44], but even more simply, visual characteristics by varying the appearance from mechanic-looking robots up to humanoid robots [10,27,41,62]. Research has shown that varying robot appearance affects whether humans attributed mental states to the robot or not (see literature on mind perception [25,62,68,70]) or the degree of other social responses [24].

Thus, human-like appearance, that is anthropomorphism operationalized through appearance, can create the feeling of social presence [1,32,35]. Social presence in turn means the "sense of being with another" [4, p. 456] and is a hint that humans actually categorize the robot or another artificial agent not as a tool, but as a social partner with whom one might interact [14]. However, social presence needs to be differentiated from physical presence. While these two concepts are different by definition (since a device or agent being physically present does not necessarily evoke the feeling of social presence if it is perceived as a tool rather than a social actor), physical presence might enhance the feeling of social presence [40].

To sum up, the CASA assumption was the starting point in studying social-psychological phenomena such as the social desirability effect in human–technology interaction using computers, smartphones and finally robots. However, in order to understand variance in effects, more fine-grained theories are needed. One key-factor especially in human–robot interaction is thus anthropomorphism that is used to explain why technology is perceived more or less as a social actor. Subsequently, different factors were identified that evoke different levels of anthropomorphism, even unconsciously, including motivational factors or design features such as a humanoid appearance. These factors serve as social cues that activate other concepts such as the feeling of social presence.

2.3 Social Desirability Bias in Human–Technology Interaction

On the basis of the theoretical background in Sect. 2.2 it can thus be argued that human users apply social norms toward technology. While the CASA assumption would expect this to be the case for all kind of technology, anthropomorphism-related theories would expect this to be higher with technology showing more social cues and thus evoke anthropomorphism and subsequently the feeling of social presence. This in turn is argued to cause humans to apply social norms.

These assumptions built the basis for research on effects of social desirable responding in human–technology interaction. While psychological research had argued for lower social desirability bias in computerized interviews compared to interviews between two humans [55], Nass et al. [48] was upon the first to study the social desirability effect under the assumption to find distortion also in human–computer interaction.

In Nass et al.'s [48] classic experiment, participants first worked with a computer on a task and had to rate the performance of the computer afterwards. For the performance rating, participants were grouped into one of three conditions: The interview was conducted by (1) the same computer, (2) a different computer, or (3) a paper-pencil questionnaire. The researchers then compared the ratings of the latter two groups to the ratings of the condition with the same computer and hypothesized that ratings should be more positive as an indicator for social desirability distortion. The interviewer-based bias would affect the condition in which the computer evaluates itself the most, and less the situations in which another computer device evaluates the interaction, since in the latter situation no "one" is offended directly. However, under assumption of the CASA theory, another computer would still be perceived as a being and thus distortion would be greater compared to the paper-pencil condition. They tested this in two experiments and found significant differences in comparison to both of the other groups in the predicted direction—for both experiments. Effect sizes were large with Cohen's d ranging from 1.07 up to 2.06 (!)—however, with extremely large confidence intervals. This findings need to be interpreted with caution as many problems that we discussed in light of the replicability crisis, can be identified here. First, the study only had a very small sample size of just $n = 10$ per group. Considering that in social-psychological studies one should expect only rather small effect sizes on average [58], this study is most likely underpowered making it more likely to be a false-positive finding. Tourangeau, Couper, and Steiger [65] sharply criticized Nass et al.'s findings because of this small sample size, the sample being a student sample, and the laboratory setting. Several other experiments [9,28,33,65] thus built upon these results. They tried to conceptually replicate the effect of social desirability distortions, but also expand the findings by using different technological devices such as virtual characters [28], websites [33,65], or smartphones [9], and by adding further concepts to explain differences such as variations in social cues [65, i.e.]. However, opposed to Nass et al.'s [48] original study, these experiments did not find such strong effects of social desirability.

In a laboratory experiment by Hoffmann et al. [28] the social desirability effect was tested by comparing the evaluation of an interaction with a male life-sized virtual agent MAX, instead of a computer, on four scales (competence, friendliness, naturalness, activity/dominance). The evaluation was examined (1) by MAX himself, (2) on a questionnaire in the same room, or (3) on a questionnaire outside the room. However, the authors found only one significant effect: Participants rated MAX's competence significantly higher compared to the questionnaire in the same room with a medium effect size of $d = .66$, and no significant difference for all the other scales (although such an effect would have been expected for all mentioned variables based on CASA). However, although sample size was larger compared to Nass et al.'s [48] study and thus power was higher, sample size for this design was rather small ($N = 63$), indicating that the study was potentially underpowered. In addition to this, several dependent variables were used leading to problems of multiple testing and the effect was much smaller than expected based on the findings of Nass et al. [48].

Very similarly to Nass et al. [48], Karr-Wisniewski and Prietula [33] tested the social desirability distortion using different computers. As an extension to the original experiment, the researchers tested different websites as additional manipulation, leading to a 2 × 2 experiment with $N = 72$ participants. However, they did not find a significant main effect for computers, contradicting the results of Nass et al. [48]. The data showed only a difference between websites with a moderate effect size of $d = .49$ but a large confidence interval of 95% CI = [.02, .96]. Again, although sample size was higher, the study could still be considered underpowered regarding the 2 × 2 between-subject design.

In a more recent study by Carolus et al. [9] smartphones were used instead of desktop computers to replicate Nass et al. [48] results of the social desirability effect in human–technology interaction. Since technology has evolved, as well as technology usage has changed since the original CASA studies, the authors call for the need of a replication with different devices. For example, smartphone usage differs from desktop computer, as most people carry a smartphone constantly with them and use it independently from time and place for different applications. Therefore, involvement with the device should also differ, leading to different social perceptions [9]. In their study $N = 108$ participants evaluated a smartphone after an interaction phase in three different evaluation groups: The evaluation was conducted using (1) the same smartphone, (2) a different smartphone, (3) by the participants own smartphone. However, evaluations did not significantly differ when evaluation was conducted by the same smartphone (with which they interacted before) or by a different smartphone ($p = .82$, with a very small effect size $d = .015$; 95% CI = [−.31; .62]), contradicting the CASA assumption. However, the difference was significant when the evaluation rating using the same smartphone for evaluation was compared to the rating using ones own smartphone ($p = 04$; $d = .57$; 95% CI = [.10; 1.04], confirming the hypothesis that participants would rate the smartphone more negatively when using their own phone for evaluation because of feeling closer to their own device and thus leading to a less distorted answer (a more honest one). Thus, results in this study were quite mixed concerning the CASA hypothesis although having higher statistical power.

Other research focused more on the controlled manipulation of social cues as a potential explanation why social desirability might differ between situations. Early psychological research shows, that the effect of computerization depends on how the interface makes the respondent feel. For example, Richman et al. [55] found that year of publication moderated the effect of computerization compared to paper-pencil questionnaires and conclude that this moderation might be due to the improvements in interface design becoming more like the paper-pencil-versions in appearance. Thus, it seems that the effect of social desirability might depend on interface design. This means computerization effects changed across the history and this difference in effects is explained by differences in design.

More directly Tourangeau et al. [65], for example, varied the social cues such as pictures or voice resulting in different administration modes, in two web experiments and one telephone study. As dependent variable, the researchers used an impression management scale in order to measure social desirability distortions. It was expected that interfaces with more social cues should result in more social desirable responses. Although using a very large sample size (more than 1000 participants in each of the three studies) their attempt to humanize the interface had low effects. They only found one significant effect across three trials with a very small effect size. A very similar study by Schmidt-Catran and Hörstermann [59] showed also no effect of a picture as social cue. Thus, the results showed little support for the effect of social cues on social desirability although (1) having large sample size and thus high power, (2) using strong humanizing cues (and thus it was expected to lead to higher effects in applying social norms), and (3) using established measurement methods.

In sum, studies on social desirability effects in human–technology interaction were not able to confirm the large effects found by Nass et al. [48] although using higher sample sizes. The CASA hypothesis was thus only partly supported. Participants showed not always a social desirable distortion in responses toward technology [9,28,33]. Additionally, the hypothesis that more social cues would lead to more social desirable behavior cannot be considered as confirmed, either [65]. In general, effect sizes varied from large to small including zero effects. However, the studies differed in many aspects such as stimulus, the amount/strength of social cues, sample diversity, contexts and measurement methods (response bias in disclosure of sensitive information, or in the evaluation of the stimulus). It is thus possible that many other influential factors could have led to different results. In addition to these differences, many problems discussed in the replicability crisis can be identified. Most of the studies only used a small sample size and thus are likely underpowered. Additionally, analyses were not always controlled for multiple-testing and used several measures [28] making the significant results less convincing. So one cannot disentangle how much variation in results is due to variation in true effects or due to noise based on variation in methodological and theoretical quality. There are more studies needed for a meta-analytic analysis. In conclusion, the social desirability effect in human–technology interaction is still unclear and needs further investigation.

3 Research Question and Hypotheses

In order to test the overall robustness of the social desirability effect in human–technology interaction and to test the social desirability effect in human–robot interaction (instead of computers, smartphones or virtual agents) it was decided to conduct another experiment. This study is clearly not a direct replication, but it tries to replicate the basic idea and can thus be considered a conceptual replication of the social desirability effect (see [67] for a definition): a short interaction with a technological device is followed by an evaluation phase in which the evaluation is conducted in different experimental groups by using the same device additionally to two or more other devices. Social desirability is thus tested by a difference in ratings between groups. Instead of computers [48], smartphones [9] or other devices, we use robots. Additionally, we study anthropomorphism and social presence as potential variables to explain the differences.

Because the true effect is believed to vary substantially based on context characteristics such as social cues, conceptual replications allow to incorporate other contexts that can be controlled for in meta-analyses or that can be discussed in reviews when a greater body of research is achieved. These in turn would then allow to estimate the variance of the effect across situations and to test theories that can account for these differences.

This goes along with the second aim to test this effect for social robotics. While the social desirability distortion was tested in situations where a computer, a smartphone or a virtual agent were used, this study tries to test the same theoretical idea in human–robot interaction. (Humanoid) robots differ fundamentally from computers, as they can display stronger social cues merely by their physical appearance or movements, or differ from virtual agents by their physical presence. Therefore, it can be assumed, that the true social desirability distortion differs from distortions in contexts depending on the interaction-device. Based on this reasoning, robots should evoke stronger feelings of social presence and thus the social desirability distortion should be even greater than in other studies such as Nass et al. [48].

The basic idea is that variation in social cues and thus variation in perceived social presence should cause variations in response patterns. Social presence was manipulated on three levels: After a short conversation (a form of interaction) with a humanoid robot, an evaluation of this conversation was conducted by (1) the same humanoid robot, (2) a different humanoid robot, and (3) a tablet computer. Note that the variation in social presence is not during the actual conversation, but the evaluation of it since social desirability distortion occurs in evaluative situations (interviewer-based bias). The evaluation was conducted via ratings including the likability of the robot leading the conversation, and the quality of the interaction. We hypothesized, that social presence should be perceived more strongly in the robot groups compared to the tablet group. This difference would be attributable to a higher anthropomorphic perception of (especially humanoid) robots compared to computers as they provide more social cues.

H 1 User ratings of social presence and anthropomorphism are higher for robots compared to a tablet computer.

3.1 Confirmatory Hypotheses

Based on this design, two effects would be expected. First, there would be a distortion toward a more socially desirable response in the condition using the same humanoid robot compared to another humanoid robot because (1) concerning social norms, it is considered to be more okay to tell another being an honest evaluation if it is not as positive, compared to telling the person directly, and (2) the longer interaction duration with robot 1 could have additionally enhanced emotional involvement and thus social pressure to answer in a polite fashion. This would follow the argument of Nass et al. [48] and others where a more positive rating was expected using the same device compared to a different one. Additionally, based on finding on anthropomorphism, it was hypothesized that the manipulation in anthropomorphic appearance and thus the feeling of social presence (with the robots causing a higher social presence compared to a tablet computer) would cause human participants to distort their responses to a more positive value (interviewer-based bias), as social presence would evoke norms of politeness. These effects should be true for both ratings, interaction quality and the likability of the conversation robot. Interaction quality refers to the qualitative evaluation of the interaction as positive or negative and the conformity of the robotic confederate's behavior to users expectations, while likability means feeling comfort being with another and the positive association with another person or object. For both ratings it is assumed to be more polite and thus socially desirable to obtain a positive evaluation (in western culture).

H 2 The likability rating of the robot is higher if the same robot is conducting the evaluation, compared to a different robot or a tablet computer.

H 3 The interaction quality rating of the conversation is higher if the same robot is conducting the evaluation, compared to a different robot or a tablet computer.

3.2 Exploratory Research Question

Additionally, the influence of other closely related constructs was tested such as the tendency to anthropomorphize (as an interindividual predisposition) and participants' general tendency to give socially desirable answers as traits of the participants, and two additional scales on participants self-reported rather situational perceptions of the devices. These additional scales on perception were aimed to measure how

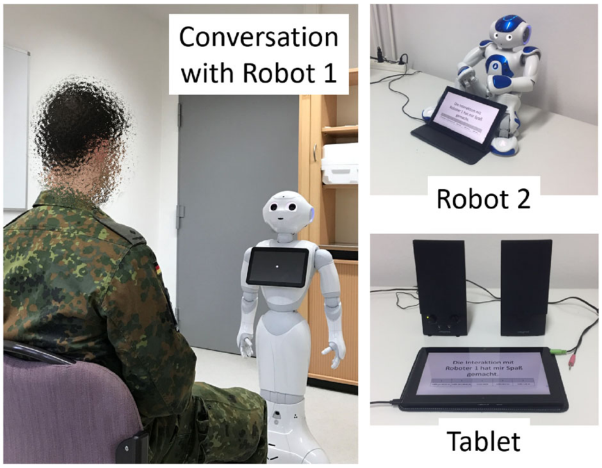

Fig. 1 Picture of robot 1 during the conversation with a participant (left side), and pictures of robot 2 and the tablet computer, both used during the evaluation (right side)

animated/alive and how responsive the devices were perceived to be. The exploratory analysis aimed to give further possibilities to explore the validity of measures and manipulation and to make it possible to test hypotheses in future attempts.

4 Method

4.1 Study Design and Sample Size Justification

The study (laboratory experiment) manipulated social presence in a between-subject design with 3 levels varying whether the evaluation was conducted by (a) the same robot with which the participants had the conversation before (robot 1), (b) a different robot (robot 2), and (c) a tablet computer (tablet). Participants were randomly assigned to one of the three conditions (see Fig. 1). As dependent variables, the evaluation included ratings of the likability of robot 1 and the quality of the interaction. The study took about 40 min and participants were rewarded with €5. The study took place in two periods in 2017 and 2018, and was examined by 6 different examiners (2 female, 4 male), which helped conducting the experiment as part of a practical training in empirical research methods.

Sample size for this experiment was calculated in a power analysis using G*Power [20]. For the calculation, a medium to large effect size was assumed of $f = .35$ and α error probability was set to .05. In order to achieve a power $(1 - \beta)$ of 80% (that is the probability of finding an effect under the assumption that there is a true effect) a minimum total sample size of $N = 84$ is needed. A total of $N = 107$ participants (36 participants in each of the robot conditions and 35 in the tablet condition) were recruited at the Bundeswehr University Munich, Germany. All participants were German speaking students with a military background. 28 Participants identified themselves as female, 79 as male, and no person as any other gender. Ages ranged from 19 to 46 ($M = 23.28$; $SD = 3.57$; $Med = 23$). Only 5 participants indicated having participated in a robot experiment before.

4.2 Dependent, Mediating and Control Variables

Dependent variables As indicators of a social desirability effect (the distortion of an evaluative rating toward a more positive value), likability of the conversation robot and the quality of the interaction were assessed. For the likability of the robot, an adapted version of the Reysen likability scale [53] including 11 items (such as "Robot 1 is likable" or "I would like to be friends with robot 1") was used. For assessing the quality of the interaction a 9-item questionnaire was used [72] (e.g. "The interaction with robot 1 was fun", "I would have loved to interact with robot 1 longer"). All items had to be rated on a 5-point Likert scale.

Additional mediating, moderating and control variables The following variables were assessed for further analysis:

1. Demographic questions including gender, age, job / major, and if the participants had taken part in a robot experiment before (for the purpose of sample description)
2. Manipulation check items (12 items including different aspects, e.g."During the evaluation (phase 2) I thought of the programmer, that could potentially see my ratings", or "I evaluated robot 1 honestly")
3. Social presence during the evaluation phase defined as "the sense of being with another" measured with a 6-item scale developed by the authors (e.g. "How much did you feel connected to the device?", see Appendix A)
4. "Animation" of the device in the evaluation phase measured with a 5-item scale developed by the authors (e.g. rating how much the device was "artificial" or "natural", see Appendix A)
5. Responsiveness of the device in the evaluation phase with a 9-item scale developed by the authors (rating how much the device was "distanced", "involved", or "absent-minded", see Appendix A)
6. Anthropomorphism (situational) in the sense as how much the device in the evaluation phase was anthropomorphized by the participant with a 6-item scale developed by the authors (e.g. how much the device "…has emotions", or "…has a personality", see Appendix A)
7. Social desirability tendency using 16 items of the SES-17 social desirability scale [61] (e.g. "I occasionally speak badly of others behind their back.")

8. Tendency to anthropomorphize (general interpersonal tendency as opposed to a situational attribution), that is participants' general tendency to attribute human-like characteristics to non-human agents, as an individual predisposition with the 15-item IDAQ scale [69] (e.g. "To what extent does a tree have a mind of its own?", "To what extent does the average robot have consciousness?")
9. Technology affinity using the 19-item TA-EG questionnaire [34] (e.g. "Electronic devices cause stress")

Note that the scales to measure "social presence", "animation", "responsiveness", "anthropomorphism (situational)" are developed by the authors and did not undergo a specific validation process. All the newly developed scales can be obtained in Appendix A. In order to measure "social presence" according to the definition of "social presence" in this article as "the sense of being with another" questions were formulated that were semantically close to this definition such as questions on how much the participant feels alone or how much they feel that "someone" else is in the room. Items of the "anthropomorphism (situational)" scale were very closely formulated to the established IDAQ questionnaire and included the ascription of intentions, free will, mind and others also found in the IDAQ [69]. "Animation" and "responsiveness" included several adjectives to describe the devices as such and were used only for exploratory analysis. We do not claim that the scales measure entirely different constructs, overlapping sub-factors of a higher-order construct or just different aspects of the same construct, since—to our knowledge – no nomological net exists that incorporates these aspects properly yet. However, we do state, that these aspects are connected to each other derived from their semantic meaning. For example, based on the reasoning in 2.3, the attribution of human-like traits such as intentions or a mind would lead to a higher feeling of social presence. Thus these concepts are closely related, but we do not indicate a more detailed factor structure.

4.3 Procedure

The experiment consisted of three parts: In the first part of the experiment, the conversation phase, participants had a short conversation with robot 1. In the second part, the evaluation phase, this conversation and robot 1 as conversation partner were evaluated accordingly. The third part consisted of various questionnaires. Before the experiment, as a cover story, participants were told that a newly developed algorithm—enabling robots to have short conversations with humans—had to be tested. They gave informed consent and agreed on data collection. Afterwards, participants were seated 88 cm in front of a robot (a humanoid "Pepper" robot from Aldebaran Robotics SAS and SoftBank Mobile Corp; called "robot 1" in the experiment) for having a short conversation with the robot in the first phase (see Fig. 1). Before the interaction in the conversation phase, the examiner left the room, so the participant was alone with the robot. During the conversation, the robot was operated via a Wizard-of-Oz (WoZ) setup. WoZ is an experimental technique in which the robots behavior including movement and speech are remotely controlled by a human operator without awareness of the participants (in this experiment the operator was seated in another room). This technique ensured full control over the robot behavior, and thus experimental rigor, while at the same time giving the participants the feeling of a realistic interaction. For the interaction, the operator used a pre-programmed set of behaviors and phrases to keep the interaction the same for every participant. The conversation was led by the robot asking questions. The outline of the conversation was largely similar for every participant since everyone was asked the same questions. A set of possible reactions was available to the operator in order to react toward users' answers to the questions and toward comments properly. Operators were trained in operating the robot prior to data collection. In the conversation, the robot introduced itself and asked some personal questions (for example on favorite color or if the participant owns a robot or has a pet), and shared short "personal" stories (for example, that it feels uncomfortable because of the bare walls in the laboratory and thus plans to paint the walls in glitter-pink). The full script of the conversation can be seen in Appendix B. Note, that the answers during this conversation were not recorded or used for further analysis—as indicators for social desirability distortion only answers in the evaluation phase were used. During the conversation, the robot moved its arms / hands (gesturing), its head and the upper body, in order to make the conversation more socially appealing. These movements include, but are not limited to, showing the palm of the hands toward the participant while asking a question to indicate their turn to answer, opening-up the arms as a welcoming gesture, moving the head from side to side as an orienting movement, slightly nodding with its head as an affirmation signaling that the participant's answer was understood, or pointing toward itself with a hand when talking about itself. The conversation was about 1.5 minutes long (this was similar for all participants because the questions required only short one-word answers such as "yes" or "no" or naming a color). After the conversation, the examiner returned and introduced the second phase of the experiment: The evaluation of the interaction. Hereby, the participants were randomly assigned to one of three conditions: in the three conditions, the evaluation was conducted (a) by the same Pepper-robot used for the conversation (robot1), (b) by a different robot (a humanoid "NAO" robot from Aldebaran Robotics SAS and SoftBank Mobile Corp; called "robot 2" in the experiment), and (c) a tablet computer (see Fig. 1). The evaluation took place in the same room, in which the conversation took place. For condition

(b) and (c), the Pepper-robot was removed from the room. Before evaluation, the examiner again left the room leaving the participant alone with the evaluation-device. An impression of the conversation with Robot 1, as well as the other two devices used in the evaluation phase (robot 2 and the tablet computer) are depicted in Fig. 1. In all of the conditions, the participants had to rate the likability of "robot 1" and the quality of the human–robot interaction as dependent variables. All items were presented visually on the devices and read aloud by the devices using a computer voice. A wizard in the other room operated all evaluation devices. Participants' ratings were given verbally and were recorded. In the conditions with the robots, the robots did not move during this phase (to ensure that the devices only differed in appearance). After this evaluation phase, the participants filled out computerized final questionnaires as listed under Sect. 4.2. For this purpose, a different computer was used (none of the devices in the evaluation phase).

5 Results

For data analysis, the open source statistic software RStudio was used [56] with packages "psych" [52], "lawstat" [21], "MASS" [66], "lm.beta" [2], and "effsize" [64].

5.1 Descriptive Statistics and Internal Consistency

Before inferential statistical analysis, internal consistency as an estimate for reliability was calculated for each scale. Single items were excluded, if exclusion led to an increase in internal consistency based on a standardized Cronbach's α.
Quality of interaction. The initial reliability of all 9 items of the scale was questionable ($\alpha = .63$). Through exclusion of 3 items, reliability could be improved to an acceptable level ($\alpha = .73$).
Likability. For the Reysen likability scale 10 of 11 items were included in the following analyses in order to improve reliability ($\alpha = .81$).
Social presence. Similarly, for the social presence scale all 6 items were included in the following analyses because of good reliability ($\alpha = .80$).
Animation. For the animation scale all 5 items were included in the following. Cronbach's α was good to excellent ($\alpha = .85$).
Responsiveness. The responsiveness scale also reached good reliability ($\alpha = .84$) and thus no item was excluded.
Anthropomorphism (situational). After exclusion of one item, the scale on how much participants anthropomorphized the robot reached excellent internal consistency ($\alpha = .91$).
Social desirability tendency. Although the authors [61] report good internal consistency, in our study, internal consistency was rather poor with $\alpha = .32$ and could only improved to $\alpha = .43$ by excluding one item. All but one items were included in the following analyses.
Anthropomorphism (tendency). The IDAQ showed good internal consistency ($\alpha = .82$) and thus none of the 15 items had been excluded.

Based on these analyses, the sum of each scale was calculated reflecting the latent construct in question to a certain extent (note that the true amount of which the real construct is reflected in the measurements is not known, it is just assumed to reflect the true construct sufficiently enough). The correlations of all variables, as well as means and standard deviations are presented in Table 1.

5.2 Manipulation Checks: Social Presence and Anthropomorphism

As manipulation check, ratings of social presence and the anthropomorphism of the evaluation-devices in each group were compared. It was hypothesized that social presence is perceived to be stronger in the robot conditions compared to the tablet condition, and that anthropomorphism ratings are higher in the robot conditions compared to the tablet (many-to-one comparisons). For each comparison, a non-parametric Brunner–Munzel test [7] was performed, and for the two comparisons ($\phi_{robot1-tablet}$ $\phi_{robot2-tablet}$) on each variable, a global α level of .05 was set. However, because of multiple testing, the local α level had to be adjusted in order to control the family-wise error rate (FWER). Therefore, the step-down Bonferroni–Holm method was used. For social presence, the difference between the evaluation using robot 1 ($M = 2.91$, $SD = .78$) and the evaluation using the tablet device ($M = 2.33$, $SD = .80$) was significant ($WB(69) = -3.28$, $p = .003$, Cohen's $d = .72$, 95% C.I. = [.23; 1.21]). In contrast, the difference between the evaluation using robot 2 ($M = 2.59$, $SD = .74$) and the evaluation using the tablet was not ($WB(65) = -1.40$, $p = .17$, Cohen's $d = .33$, 95% C.I. = [−.15; .81]).

None of the comparisons for the anthropomorphism ratings were significant: neither the comparison between the group with robot 1 ($M = 1.98$, $SD = .89$) and the tablet group ($M = 1.57$, $SD = .68$) ($WB(64) = -1.75$, $p = .17$, Cohen's $d = .52$, 95% C.I. = [.04; 1.00]). Nor between the group with robot 2 ($M = 1.59$, $SD = .73$) and the tablet group ($WB(66) = .20$, $p = .84$, Cohen's $d = .03$, 95% -C.I. = [−.44; .51]). Thus, hypothesis H1 has to be rejected, as only one difference turned out to be significant (difference in social presence between robot 1 and tablet computer).

5.3 Confirmatory Results: Group Differences

The confirmatory analysis included the comparison (many-to-one comparison) of the likability of robot 1 and on the evaluation of the quality of the interaction between

Table 1 Correlations, means (*M*) and standard deviations (*SD*) for all variables (sum scores)

	Quality of interaction	Likability	Social presence	Animation	Res-ponsiveness	Anthropomorphism	Social desirability	IDAQ
Quality of interaction	1							
Likability	**0.48**	1						
Social presence	0.11	0.28	1					
Animation	0.09	**0.39**	**0.69**	1				
Responsiveness	−0.22	−0.25	**−0.60**	**−0.59**	1			
Anthropomorphism	0.19	0.30	**.57**	**0.50**	**−0.44**	1		
Social desirability	0.16	0.19	0.01	−0.01	0.02	0.01	1	
IDAQ	0.22	0.07	0.06	−0.01	−0.07	−0.23	0.01	1
M	4.21	2.97	2.61	2.02	2.67	1.71	0.60	2.85
SD	0.50	0.65	0.80	0.78	0.77	0.79	0.13	0.72

All significant correlations (with a global a level of .05) are highlighted in bold. Note that the local alpha-levels were adjusted using the Bonferroni-procedure

1. the condition, in which robot 1 conducted the evaluation by itself (“robot1”), with
2. the condition, in which robot 2 conducted the evaluation (“robot2”; $\phi_{robot1-robot2}$), and
3. the condition, in which the tablet computer conducted the evaluation (“tablet”; $\phi_{robot1-tablet}$).

As before, for each comparison a non-parametric Brunner–Munzel test [7] was performed, and for the two comparisons ($\phi_{robot1-robot2}$ $\phi_{robot1-tablet}$) on each variable, a global α level of .05 was set. The local α level was adjusted using the step-down Bonferroni–Holm method in order to control the FWER.

For the likability rating, neither the difference between “robot1” ($M = 3.05$, $SD = .74$) and “robot2” ($M = 2.78$, $SD = .58$) was significant ($WB(65) = -1.74$, $p = .17$, Cohen’s $d = .41$, 95% C.I. = [−.07; .88]), nor the difference between “robot1” and “tablet” ($M = 3.09$, $SD = .60$) ($WB(67) = .13$, $p = .90$, Cohen’s $d = -.06$, 95% C.I. = [−.54; .41]). Thus, hypothesis (H2) had to be rejected: no difference between the groups was found for the likability ratings. For the rating on the quality of the interaction, results were similar: No significant difference emerged. Neither between “robot1” ($M = 4.16$, $SD = .54$) and “robot2” ($M = 4.13$, $SD = .45$) ($WB(66) = -.63$, $p = .53$, Cohen’s $d = .07$, 95% C.I. = [−.40; .54]), nor between “robot1” and “tablet” ($M = 4.36$, $SD = .48$) ($WB(67) = 1.68$, $p = .19$, Cohen’s $d = -.38$, 95% C.I. = [−.86; .10]). The hypothesis for the quality of interaction rating (H3) was rejected: no difference between the groups was found. It should also be noted, that besides non-significant differences, effect sizes are also rather low and the 95% confidence intervals are large, thus indicating a considerable amount of uncertainty of the estimation.[2]

5.4 Exploratory Results: Step-Wise Regression

The reader should note that the results reported in this section are of exploratory nature and data driven. No hypothesis had been formulated in advance. Additionally, especially “animation” and “responsiveness” were newly developed scales and did not undergo a rigorous validation process. Results should be interpreted with caution because the following results can also be attributed to random variation. Thus, results here should not be considered as “findings”, but rather as input for new hypotheses that have to be investigated in further experiments.

For the exploratory analyses, step-wise linear regressions with a forward selection of variables were conducted. This data-driven approach in several steps results in the best-fitting model. As a stopping rule of selection, the Akaike information criterion (AIC) was chosen. The quality of the interaction and likability were used as dependent variables. As predictors, animation, responsiveness, anthropomorphism ratings of the robot (situational), social desirability tendency and the tendency to anthropomorphize (IDAQ) were included in the procedure. For likability as dependent variable, the step-wise

[2] Grouping together the manipulation checks and the confirmatory analyses, 6 tests were performed overall. FWER was controlled for each variable separately by adjusting the p value using the Bonferroni–Holm step-down method. In these analyses presented above, FWER is not controlled considering all 6 tests. However, by adjusting the p values with respect to all 6 tests, the difference in social presence ratings (between the group with robot 1 and the tablet) remained significant with an adjusted value of $p = .01$. Since all of the other comparisons were non-significant before, this was of course true after additional adjustment—needless to say, that values were even larger (range = [.59, 1.00]).

procedure resulted in a model including animation ($\beta = .35$) and the tendency to answer socially desirable ($\beta = .19$) as predictors. However, the overall model describes the variation rather poorly ($R^2 = .16$). For the ratings of the quality of the interaction, the step-wise procedure resulted in a model including anthropomorphism (situational) ($\beta = .23$) and the tendency for social desirability ($\beta = .19$), too. However, the coefficient of determination being close to zero ($R^2 = .09$), indicates that the model describes the variation in ratings of the quality of interaction even worse. Considering low values for R-squared and low β values of the predictors, the models might be the best fitting ones based on AIC, but cannot be considered as overall acceptable in predicting the dependent variables in question.

6 Dicussion

In this study, we tested the social desirability effect in human–robot interaction. In three groups, participants evaluated the likability of a social robot and the quality of the interaction, with the interview being conducted (a) by the same robot itself, (b) by a different robot, and (c) by a tablet computer. This study design is very similar to other studies that used various different devices such as computers [48], websites [33], virtual agents [28], or smartphones [9] and thus is considered being a conceptual replication of such an effect. Additionally, it was expected that the robots evoke a higher level of anthropomorphism and more sense of social presence compared to the tablet because of displaying more social cues. According to the CASA assumption and more specifically literature on anthropomorphic design, ratings of likability and quality of interaction were expected to be higher if the robot conducted the interview itself, compared to a different robot or a tablet computer because the higher feeling of social presence (due to more visual social cues and higher emotional involvement) should lead to more socially desirable responding.

The manipulation in this study was only partly successful, since only the difference in social presence between the conversation-robot (robot 1) and the tablet was significant, but not between the other robot (robot 2) and the tablet. In this study, social presence and anthropomorphism were not as easily induced by robot appearance as other research had indicated [47]. A similar failure to induce a significant amount of anthropomorphism and thus a sense of social presence with social cues was also reported elsewhere by finding weak effects [65] or non-significant results [1, $p > .05$]. However, this shouldn't make a difference considering a social desirability bias according to the CASA theory. The manipulation check of anthropomorphism and social presence were explicit measures. According to Reeves and Nass [51], social norms are applied automatically even without participants explicitly reporting higher anthropomorphism. Similarly, the dual-process theories [74] could explain how—even without finding a difference in explicit measures—implicit processes could lead to an automatic activation of these concepts and thus still lead to (unconscious) social desirability bias. Thus, social desirability distortion would still be expected, even without explicit anthropomorphism or conscious perception of social presence.

Yet, the data of this experiment did not support this claim either. The ratings of likability and the quality of the interaction did not differ significantly, effect sizes were small and confidence intervals were large. The CASA theory was not supported by this experiment. Results—the effect size magnitude—of Nass et al. [48] were not replicated, although (1) sample size and thus statistical power was higher, and (2) social cues were expected to be stronger in robots compared to computers. However, the reader should note that the aphorism "absence of evidence is not evidence of absence" also should be considered here. The absence of evidence can have various reasons of which is just one the absence of a true effect:

1. Of course one reason is that there is no true social desirability effect in human–technology interaction as suggested by the CASA hypothesis, or at least it is negligibly small. Thus the hypothesis that humans would distort their responses in a social desirable direction because technology invokes a sense of social presence is correctly rejected based on the results of this experiment. When reviewing the literature, the effect seems to vary vastly and did not reliably reveal such an effect. And indeed, taken into consideration that (1) if there is no true effect, the p values would be distributed equally (that means small p values are as likely as large p values), and (2) publication bias favors positive results, the mixed picture in the literature could exactly stem from just this: no true effect.
2. Another explanation of the mixed picture in the literature and the absence of evidence in this experiment can be that there is a true effect, it just varies a lot across situations—including situations where the true effect is very small or even close to zero. A variation in the literature thus is caused by the variation of true effect sizes. This variation can for example be due to variation in suggested factors such as social cues of the technology-design. However, there are plenty potential influential factors that lead to a variation in effect sizes besides design features. The more advanced three-factor theory by Epley et al. [15] describes other factors that can account for this variation. But there might be additional factors, that have not been explored very much yet. More theory is thus needed. For example the social desirability effect in a situation as in this study could have been smaller as

in others because of certain-yet-not-known aspects of the sample or certain characteristics of the interaction. One aspect is the very specific sample of mainly male, rather young military students that could behave differently than more diverse samples. Other aspects concern the experimental setup. For example, participants could have rated the researchers' ability to create a good conversation algorithm and in doing so they thought about the "human behind the machine" instead of the robot directly and thus it was more a social interaction with the researchers. However, participants tended to think little of the programmer ($M = 2.54$, $SD = 1.66$ using a 5-point Likert scale) and only moderately about "the human behind the machine" ($M = 3.30$, $SD = 1.72$ using a 5-point Likert scale). Another aspect is the conversation with the robot. The conversation was about 1.5 min long and included topics such as pets, favorite color or short anecdotes meant to be humorous (painting the laboratory walls in glitter-pink). However, there is possibility, that the conversation was too short in duration and the topics were not critical enough in order to evoke the need of socially desirable distortion. It is also possible that the conversation was actually fine and participants did not need to distort the response in a positive direction, because they really meant it. This would result in a ceiling effect. Indeed, the rating of the quality of the interaction was quite high with a small variation. In contrast, likability was only moderately high and showed larger variation—based on the given theory one would thus have expected a social desirability effect at least in this rating. The criticality of the interaction is a crucial moderating factor for social desirability bias to occur in the first place. Even though the likability-rating was only moderate and we thus would expect the situation to be critical enough for the bias to be non-zero, the social desirability effect could still be so small, that sample size and power were just too low to detect the bias. In conclusion, with a more controversial conversation or some small errors, maybe participants would have shown larger response bias (because of a higher need to distort responses in a socially desirable direction)—large enough to turn out significant, or under the assumption of small effects even larger sample sizes are needed. Reeves and Nass [51] proposed Grice's maxims—four conversational rules based on high quality, appropriate quantity, high relevance, and clarity—in order to achieve polite interaction [26]. Thus, a violation of one of the rules as a manipulation could be used to achieve a more impolite conversation and thus more bias.

3. A third concern are the appropriateness of manipulation and validity issues. The manipulation of appearance and social involvement was expected to induce enough social cues in order to trigger the application of social behavior toward the robot. However, it is possible that manipulation of appearance is not enough in order to obtain effects that are large enough. A possible solution would be to install more social cues in the robot condition, resulting in a larger difference between groups regarding social presence. For example, the conversation robot could be given a name and when evaluating itself, it could actually refer to itself as "me" instead of "robot 1" as in the other conditions (e.g. "How much do you like me?"). This could lead to higher anthropomorphism, higher social presence and therefore to higher social desirability biases. Furthermore it is also possible that the dependent variables are not valid in this context. For example, the quality of interaction scale did not undergo an extensive validation process. And although as a measure of likability a validated scale was used [53, Reysen likability scale], it is unclear whether this measure is similarly appropriate for the evaluation of robots. Since the scale had been developed for human–human evaluation, it is questionable if it can also be applied to human–robot evaluation. It probably has to be adjusted. For example, the original scale includes items such as "I would like this person as a coworker" (Here: "I would like robot 1 as a coworker"). However, since the Pepper-robot was developed for purposes of social interaction and not with the purpose to be used in all kinds of work environments, this item might not reflect likability as much as when evaluating another human adult. Moreover, the evaluation might be context-dependent. For instance, the sample consisted of students with a military background, so a Pepper-robot is probably not considered a good coworker in military settings, even if the participants liked it. This might result in a response pattern that is invalid for measuring likability as a construct since variation in this item does not reflect variation of likability. Thus, scales might have to be adapted to the specific context of the evaluation and validated accordingly. This means that even if there is a true effect and the size of this effect is sufficiently high in this situation, lacking validity of the measurement methods might have had the effect that the true social desirability effect was not reflected by these measures.
4. A fourth cause of variation and null-effects can be the very same reasons discussed in context of the replicability crisis. Even if (1) there is a true effect size, (2) this true effect size is sufficiently large in the situation of an experiment, and (3) the measurement methods are appropriate and the manipulations are valid, there is still a variation that can lead to an effect not being detected: Statistical-methodological weaknesses such as small sample sizes and thus under-powered study designs and questionable research practices. We used a medium to large effect size, in order to calculate the number of participants needed to achieve a power of 80% – that is the probability to detect

an effect under the assumption there is a true effect, or in other words the probability of correctly rejecting the null hypothesis given that the alternative hypothesis is true. However, meta-science clearly suggests that for social psychology only medium, if not rather small effect sizes have to be expected on average [54,58]. For example, an overview of effect sizes in social psychology as a subfield showed that the effect size typically is about $r = .21$. According to this, the effect size that was expected in advance, was too optimistic. This can also explain, why other studies in human–robot interaction failed to find effects when studying social-psychological phenomena [31]. Effect sizes may be simply overestimated [58]. Similarly, studies on social desirability in human–computer interaction expected rather large effects, but data did not always show such large effects [65]. That means, many studies based on this assumption [28], including the experiment reported in this article, are therefore underpowered. Hence, these studies have a low probability of finding a true effect and thus produce more false positive outcomes [8]. Moreover, even when finding a significant effect, it is likely that the estimate of the magnitude of the effect is inflated and thus overestimated—a phenomenon called the winner's curse [8]. This means that finding no effect of social presence on socially desirable responding, can simply be due to the sample size being too small (under the more realistic assumption of a medium to small effect size). For example, calculating the sample size using the same parameters, but an effect size of $r = .21$ as mentioned above, a total of $N = 222$ participants would have been needed in this 3-group between-subject design ($n = 74$ per group). Furthermore questionable research practices as mentioned in Sect. 2.1 worsen the problem even more, but had been widely used in the psychological literature. However, in this experiment we reported all analyses—both confirmatory and exploratory ones—accordingly (there had not been further analyses as the ones reported here), controlled for FWER, did not (secretly) delete single participants or whole groups of participants (as we had no legitimization), and did not control for different variables a posteriori. We therefore tried to prevent the (subconscious) use of questionable research practices to the best of our knowledge.

In sum, many concerns were raised that are not only problematic for this study, but also for other studies in social robotics. As other studies before [see 9,65], results indicate limited support for the CASA assumption. Of course this study reported here had many limitations that can account for not finding an effect such as a rather specific sample size, the conversation not being critical enough or too few social cues. However, taken together (1) the overall mixed picture including the results of the experiment reported in this article and other studies in the literature, and (2) the overall low power in social robotics experiments [see for example 39], chances are, that the social desirability effect as suggested by Nass et al. [48] is overesitimated. Note that this does not mean that the theory does not hold, but it has to be considered that the effect size might be much smaller on average and that effect sizes vary to a wide degree across situations and persons. This could mean that such results are real, but oftentimes practically not meaningful because of small effect sizes. However, factors need to be identified that lead to strong effects and define situations in which such a social desirability effect needs to be considered in evaluation design. For example, the results from the exploratory analysis show influence of animation and of a general tendency to answer socially desirable. Social desirability tendency and also the perception of animation seemed to influence the evaluation in general. Maybe in our study, interindividual differences of the participants had thus more effect on variance than the actual interaction design. However, the effect sizes are negligible small and as noted before, these results were solely data-driven and thus interpreted with caution. However, instead of simply assuming that computers or robots are treated as social actors, more careful reflection is needed on why and to what degree this can be expected in a given situation. Therefore, theories such as the three-factor theory of anthropomorphism [15], contextual factors and interindividual differences need to be studied more extensively. Sample sizes should then be justified based on these theoretical considerations and effect sizes should rather expected to be small than optimistically large. A social desirability effect in human–robot interaction would have very practical implications especially for the evaluation of social robots. If the social desirability effect is true it would be recommended that a robot or its service should not be evaluated by the very same robot, but by a different device instead, in order to achieve an unbiased feedback. However, the evaluation of a robot by itself after a service would be much easier as no other device is needed and thus saves costs, effort and time. The results of this experiment and other results in the literature show only mixed support for social desirability bias. It was argued that the effect could be overestimated and in fact could be rather small for many situations and thus a robot evaluating itself can provide still acceptable estimates. In such situations, a bias that small could be practically negligible and a correction would not be worth the extra costs and effort of installing another device. Although we did not find effects of social cues, following anthropomorphism theories, an easier strategy to reduce social desirability bias (at least a bit) could be to change the interaction into a different less socially appealing more mechanic mode. For example, while during the service a social robot could appear more as a social actor through gesturing, friendly voice style and natural-appearing conversation (i.e. making jokes, referring to itself as "me"), in the evaluation mode these social cues could be reduced in making the robot more mechanic by not showing gestures, mechanic voice and more neutral language (i.e. referring to itself in third person as "robot")—so even if it is the same robot it appears more distanced maybe even almost like a different device. To sum up, we want to stress once more to not overgeneralize the results in this study, because finding no effect, is not evidence that the social desirability effect does not exist or is only small. And thus, if it is possible and if it does not take much effort, we would recommend to stick with the more safer conclusion to expect a social desirability effect in HRI and use different devices for service evaluation—especially if the accuracy of the evaluation is of high importance. On the other side taken together the results of this experiment and other experiments in the literature the social desirability effect seems not to be very large—if not even small – in many situations. Therefore, evaluations by the technology itself (which might be more convenient) can provide acceptable estimates.

Acknowledgements Open Access funding provided by Projekt DEAL. This research was funded by the Bundeswehr University Munich. The authors like to thank Philipp Becker, Julian Fleskes, Simon Freudling, Julia Jahn, Carolin Stenglein, and Eric Wenzel for their help in conducting the study.

Funding This research was funded by the Bundeswehr University Munich.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval We did not seek ethical approval from the ethics committee, as our study does not fulfill criteria where ethical approval is required in Germany. The reported study is non-invasive and non-clinical research on human subjects and subjects were informed about the purpose, aim, and risks of the studies and that risks are reasonably low. A coverstory was used, whereby participants were told before the experiment started that the purpose of the study was to test a conversation algorithm. Immediately after the experiment, participants were fully debriefed and informed of the actual focus of the study (social desirability bias), as well as the Wizard-of-Oz design. Participants were informed before the experiment that they are free to quit participation and data collection at any time during the experiment without any negative consequences—participation was thus voluntary. Our institute holds high data protection standards conforming to the current EU data privacy regulations. All data were collected in strictly pseudonymous form during the experiment and anonymized for data analysis later. Participants cannot be identified based on the data, since no personal data such as names and no pictures, videos or sound were stored. Every participant gave informed consent and agreed on being able and willing to participate in the experiment and data collection process voluntarily. Participant privacy is a key value that our university has committed itself to uphold.

Availability of data and material Upon request.

Code availability For data analysis, the open-source software RStudio was used. The R-code used for data analysis is available.

Appendix A

Self-developed scales for anthropomorphism (situational), social presence, animation and responsiveness. All items were rated on a 5-point Likert scale ranging from "not al all" to "very much".

Anthropomorphism (situational)

How much do you agree with these statements? Please answer spontaneously—we are interested in your opinion, therefore there are no correct or incorrect answers.

[*Device name*] had…

1. …intentions.
2. …consciousness.
3. …emotions.
4. …personality.
5. …free will.
6. …a mind of its own.

Social presence

How much do you agree with these statements? Please answer spontaneously—we are interested in your opinion, therefore there are no correct or incorrect answers.

1. How much did you feel like interacting with an intelligent being?
2. How much did you feel like being alone?
3. How much did you feel connected to [*device name*]?
4. How much did you feel that [*device name*] was sensitive to you?
5. With [*device name*], how much did you feel that someone else was in the room with you?
6. How much did you feel [*device name*] was with you?

Animation

How much do you agree with these statements? Please answer spontaneously—we are interested in your opinion, therefore there are no correct or incorrect answers.

[*Device name*] was…

1. …artificial.
2. …technical.
3. …natural.
4. …lively.
5. …human-like.

Responsiveness

How much do you agree with these statements? Please answer spontaneously—we are interested in your opinion, therefore there are no correct or incorrect answers.

[*Device name*] was…

1. …active.
2. …apathetic.
3. …absent-minded.
4. …real.
5. …distanced.
6. …interested.
7. …bored.
8. …involved.
9. …impersonal.

Appendix B

Transcript of the dialogue.

[*The situation starts with the Pepper robot in a bent over position to make a sleeping-like impression. As the interaction starts the robot awakes by going from a bent over position to an upright position. The robot looks a round by moving its body and head to the left and right before orienting and looking to the participant.*]

Robot: Welcome to the experiment! Nice to have you here! I'd like to have a conversation with you. My name is Robot 1. What is your name?

Participant: [*answers*]

Robot: Great. One day, it will be my destiny to help people and keep them company. Do you have a robot at home?

Participant: [*answers*]

Robot: That's too bad [*it was assumed that the participants' answer is "no"*]. I'm still new at the institute and only a few months old. I wonder how old you are?

Participant: [*answers*]

Robot: I see. I'm not so comfortable in this room. [*looks around the room by moving his head*] It is a little bare, isn't it? Color would be good for it. What's your favorite color?

Participant: [*answers*]

Robot: That's a nice colour too! My favorite color is glitter-pink [*it was assumed that the participants would not choose glitter-pink as favorite color*]. I wish I had a glitter-pink fish. Do you have any pets?

Participant: [*answers*]

Robot: Okay, how time flies. The experimenter should be back in a minute. Thank you for the nice, albeit brief, conversation. It was great to meet you! [*The robot is bending over to make the impression it is on stand-by*]

[*End of conversation.*]

References

1. Appel J, von der Pütten A, Krämer NC, Gratch J (2012) Does humanity matter? analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. Adv Hum Comput Interact 2:1–10. https://doi.org/10.1155/2012/324694
2. Behrendt S (2014) lm.beta: add standardized regression coefficients to lm-objects. https://CRAN.R-project.org/package=lm.beta
3. Belhassein K, Buisan G, Clodic A, Alami R (2019) Towards methodological principles for user studies in human–robot interaction. In: Test methods and metrics for effective HRI in collaborative human–robot teams workshop, ACM/IEEE international conference on human–robot interaction, Daegu
4. Biocca F, Harms C, Burgoon JK (2003) Toward a more robust theory and measure of social presence: review and suggested criteria. Presence Teleoper Virtual Environ 12(5):456–480. https://doi.org/10.1162/105474603322761270
5. Borsboom D (2008) Latent variable theory. Meas Interdiscip Res Perspect 6(1–2):25–53. https://doi.org/10.1080/15366360802035497
6. Borsboom D, Mellenbergh GJ, van Heerden J (2004) The concept of validity. Psychol Rev 111(4):1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061
7. Brunner E, Munzel U (2000) The nonparametric behrens-fisher problem: asymptotic theory and a small-sample approximation. Biomet J 42(1):17–25. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U
8. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14(5):365–376. https://doi.org/10.1038/nrn3475

9. Carolus A, Schmidt C, Schneider F, Mayr J, Muench R (2018) Are people polite to smartphones? In: Kurosu M (ed) Human–computer interaction. Interaction in context. HCI 2018. Lecture Notes in Computer Science, vol 10902, Springer, Cham, pp 500–511. https://doi.org/10.1007/978-3-319-91244-8_39
10. Carpinella CM, Wyman AB, Perez MA, Stroessner SJ (2017) The robotic social attributes scale (rosas). In: HRI '17: Proceedings of the 2017 ACM/IEEE international conference on human–robot interaction. HRI'17, IEEE, Piscataway, pp 254–262. https://doi.org/10.1145/2909824.3020208
11. Cialdini RB, Trost MR (1998) Social influence: social norms, conformity and compliance. In: Gilbert DT, Fiske ST, Lindzey G (eds) The handbook of social psychology. Oxford University Press, New York, pp 151–192
12. Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52(4):281–302. https://doi.org/10.1037/h0040957
13. Cunningham WA, Zelazo PD, Packer DJ, van Bavel JJ (2007) The iterative reprocessing model: a multilevel framework for attitudes and evaluation. Soc Cognit 25(5):736–760. https://doi.org/10.1521/soco.2007.25.5.736
14. Damiano L, Dumouchel P (2018) Anthropomorphism in human-robot co-evolution. Front Psychol. https://doi.org/10.3389/fpsyg.2018.00468
15. Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114(4):864–886. https://doi.org/10.1037/0033-295X.114.4.864
16. Epley N, Waytz A, Akalis S, Cacioppo JT (2008) When we need a human: motivational determinants of anthropomorphism. Soc Cognit 26(2):143–155. https://doi.org/10.1521/soco.2008.26.2.143
17. Eyssel F, Kuchenbrandt D (2012) Social categorization of social robots: anthropomorphism as a function of robot group membership. Br J Soc Psychol 51(4):724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x
18. Eyssel F, Reich N (2013) Loneliness makes the heart grow fonder (of robots)—on the effects of loneliness on psychological anthropomorphism. In: 8th ACM/IEEE international conference on human–robot interaction (HRI), pp 121–122. https://doi.org/10.1109/HRI.2013.6483531
19. Eyssel F, Kuchenbrandt D, Bobinger S, de Ruiter L, Hegel F (2012) 'If you sound like me, you must be more human': on the interplay of robot and user features on human–robot acceptance and anthropomorphism. In: 2012 7th ACM/IEEE international conference on human–robot interaction (HRI), Boston, pp 125–126. https://doi.org/10.1145/2157689.2157717
20. Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39(2):175–191. https://doi.org/10.3758/BF03193146
21. Gastwirth JL, Gel YR, Hui W, Lyubchich V, Miao W, Noguchi K (2017) lawstat: tools for biostatistics, public policy, and law. https://CRAN.R-project.org/package=lawstat
22. Gawronski B, Bodenhausen GV (2007) Unraveling the processes underlying evaluation: attitudes from the perspective of the ape model. Soc Cognit 25(5):687–717. https://doi.org/10.1521/soco.2007.25.5.687
23. Gawronski B, Creighton LA (2013) Dual process theories. In: The Oxford (ed) Carlston D. Oxford University Press, Handbook of social cognition, pp 282–312
24. Gong L (2008) How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. Comput Hum Behav 24(4):1494–1509. https://doi.org/10.1016/j.chb.2007.05.007
25. Gray HM, Gray K, Wegner DM (2007) Dimensions of mind perception. Science 315(5812):619. https://doi.org/10.1126/science.1134475
26. Grice HP (1975) Logic and conversation. In: Cole P, Morgan JJ (eds) Syntax and semantics. Academic Press, New York, pp 41–58
27. Hegel F, Lohse M, Wrede B (2009) Effects of visual appearance on the attribution of applications in social robotics. In: RO-MAN 2009—the 18th IEEE international symposium on robot and human interactive communication, pp 64–71. https://doi.org/10.1109/ROMAN.2009.5326340
28. Hoffmann L, Krämer NC, Lam-chi A, Kopp S (2009) Media equation revisited: do users show polite reactions towards an embodied agent? In: Ruttkay Z, Kipp M, Nijholt A, Vilhjálmsson HH (eds) Intelligent virtual agents, lecture notes in computer science, vol 5773, Springer, Berlin, pp 159–165. https://doi.org/10.1007/978-3-642-04380-2_19
29. Hussey I, Hughes S (2018) Hidden invalidity among fifteen commonly used measures in social and personality psychology. https://doi.org/10.31234/osf.io/7rbfp
30. Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124
31. Irfan B, Kennedy J, Lemaignan S, Papadopoulos F, Senft E, Belpaeme T (2018) Social psychology and human–robot interaction: an uneasy marriage. In: HRI '18 companion of the 2018 ACM/IEEE international conference on human–robot interaction, pp 13–20. https://doi.org/10.1145/3173386.3173389
32. Jung Y, Lee KM (2004) Effects of physical embodiment on social presence of social robots. In: Proceedings of presence, pp 80–87
33. Karr-Wisniewski P, Prietula M (2010) Casa, wasa, and the dimensions of us. Comput Hum Behav 26(6):1761–1771. https://doi.org/10.1016/j.chb.2010.07.003
34. Karrer K, Glaser C, Clemens C, Bruder C (2009) Technikaffinität erfassen–der fragebogen ta-eg. Der Mensch im Mittelpunkt technischer Systeme 8:196–201
35. Kim Y, Sundar SS (2012) Anthropomorphism of computers: is it mindful or mindless? Comput Hum Behav 28(1):241–250. https://doi.org/10.1016/j.chb.2011.09.006
36. Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, Bernstein MJ, Bocian K, Brandt MJ, Brooks B, Brumbaugh CC, Cemalcilar Z, Chandler J, Cheong W, Davis WE, Devos T, Eisner M, Frankowska N, Furrow D, Galliani EM, Hasselman F, Hicks JA, Hovermale JF, Hunt SJ, Huntsinger JR, IJzerman H, John MS, Joy-Gaba JA, Barry Kappes H, Krueger LE, Kurtz J, Levitan CA, Mallett RK, Morris WL, Nelson AJ, Nier JA, Packard G, Pilati R, Rutchick AM, Schmidt K, Skorinko JL, Smith R, Steiner TG, Storbeck J, van Swol LM, Thompson D, van 't Veer AE, Ann Vaughn L, Vranka M, Wichman AL, Woodzicka JA, Nosek BA, (2014) Investigating variation in replicability. Soc Psychol 45(3):142–152. https://doi.org/10.1027/1864-9335/a000178
37. Lakatos I (1976) Falsification and the methodology of scientific research programmes. In: Harding SG (ed) Can theories be refuted? vol 3, D. Reidel Publishing Company, Dordrecht, pp 205–259. https://doi.org/10.1007/978-94-010-1863-0_14
38. Lakens D, Evers ERK (2014) Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. Perspect Psychol Sci 9(3):278–292. https://doi.org/10.1177/1745691614528520
39. Leichtmann B, 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. https://doi.org/10.1016/j.jenvp.2019.101386
40. Li J (2015) The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int J Hum Comput Stud 77:23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001
41. Martini MC, Buzzell GA, Wiese E (2015) Agent appearance modulates mind attribution and social attention in human–robot interaction. In: Tapus A, André E, Martin JC, Ferland F, Ammi M

(eds) Social robotics. ICSR 2015, Lecture notes in computer science, vol 9388, Springer, Cham, pp 431–439. https://doi.org/10.1007/978-3-319-25554-5_43

42. Maxwell SE (2004) The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychol Methods 9(2):147–163. https://doi.org/10.1037/1082-989X.9.2.147
43. Millham J, Kellogg RW (1980) Need for social approval: impression management or self-deception? J Res Personal 14(4):445–457. https://doi.org/10.1016/0092-6566(80)90003-3
44. Morewedge CK, Preston J, Wegner DM (2007) Timescale bias in the attribution of mind. J Personal Soc Psychol 93(1):1–11. https://doi.org/10.1037/0022-3514.93.1.1
45. Mumm J, Mutlu B (2011) Human–robot proxemics: physical and psychological distancing in human–robot interaction. In: HRI '11 Proceedings of the 6th international conference on human–robot interaction, pp 331–338. https://doi.org/10.1145/1957656.1957786
46. Muthukrishna M, Henrich J (2019) A problem in theory. Nat Hum Behav 3(3):221–229. https://doi.org/10.1038/s41562-018-0522-1
47. Nass C, Moon Y (2000) Machines and mindlessness: social responses to computers. J Soc Issues 56:81–103. https://doi.org/10.1111/0022-4537.00153
48. Nass C, Moon Y, Carney P (1999) Are people polite to computers? Responses to computer-based interviewing systems1. J Appl Soc Psychol 29(5):1093–1109. https://doi.org/10.1111/j.1559-1816.1999.tb00142.x
49. Nederhof AJ (1985) Methods of coping with social desirability bias: a review. Eur J Soc Psychol 15(3):263–280. https://doi.org/10.1002/ejsp.2420150303
50. Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science (New York, NY) 349(6251):aac4716. https://doi.org/10.1126/science.aac4716
51. Reeves B, Nass CI (1997) The media equation: how people treat computers, televisions, and new media like real people and places, 2nd edn. CSLI Publ, Stanford, Calif
52. Revelle W (2018) Psych: Procedures for psychological, psychometric, and personality research. https://CRAN.R-project.org/package=psych
53. Reysen S (2005) Construction of a new scale: the reysen likability scale. Soc Behav Personal Int J 33(2):201–208. https://doi.org/10.2224/sbp.2005.33.2.201
54. Richard FD, Bond CF, Stokes-Zoota JJ (2003) One hundred years of social psychology quantitatively described. Rev Gen Psychol 7(4):331–363. https://doi.org/10.1037/1089-2680.7.4.331
55. Richman WL, Kiesler S, Weisband S, Drasgow F (1999) A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. J Appl Psychol 84(5):754–775. https://doi.org/10.1037/0021-9010.84.5.754
56. RStudio Team (2016) Rstudio: integrated development environment for r. http://www.rstudio.com/
57. Ruijten PAM, Haans A, Ham J, Midden CJH (2019) Perceived human-likeness of social robots: testing the rasch model as a method for measuring anthropomorphism. International Journal of Social Robotics 11(3):477–494. https://doi.org/10.1007/s12369-019-00516-z
58. Schäfer T, Schwarz MA (2019) The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. Front Psychol 10:813. https://doi.org/10.3389/fpsyg.2019.00813
59. Schmidt-Catran AW, Hörstermann K (2013) Soziale präsenz in onlineumfragen: 36 pages / methods, data, no 3 (2013) / methods, data, analyses, vol 7, no 3 (2013). methoden, daten, analyse 7(3):397–432, 10.12758/mda.2013.019
60. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science 22(11):1359–1366. https://doi.org/10.1177/0956797611417632
61. Stöber J (1999) Die soziale-erwünschtheits-skala-17 (ses-17): Entwicklung und erste befunde zu reliabilität und validität. Diagnostica 45(4):173–177. https://doi.org/10.1026//0012-1924.45.4.173
62. Stroessner SJ, Benitez J (2019) The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. Int J Soc Robot 11(2):305–315. https://doi.org/10.1007/s12369-018-0502-7
63. Świątkowski W, Dompnier B (2017) Replicability crisis in social psychology: looking at the past to find new pathways for the future. Int Rev Soc Psychol 30(1):111–124. https://doi.org/10.5334/irsp.66
64. Torchiano M (2019) effsize: efficient effect size computation. https://CRAN.R-project.org/package=effsize
65. Tourangeau R, Couper MP, Steiger DM (2003) Humanizing self-administered surveys: experiments on social presence in web and ivr surveys. Comput Hum Behav 19(1):1–24. https://doi.org/10.1016/S0747-5632(02)00032-8
66. Venables WN, Ripley BD (2002) Modern applied statistics with s. http://www.stats.ox.ac.uk/pub/MASS4
67. Ward M, Kemp S (2019) The probability of conceptual replication and the variability of effect size. Methods Psychol. https://doi.org/10.1016/j.metip.2019.100002
68. Waytz A, Cacioppo J, Epley N (2010) Who sees human? The stability and importance of individual differences in anthropomorphism. Perspect Psychol Sci 5(3):219–232. https://doi.org/10.1177/1745691610369336
69. Waytz A, Gray K, Epley N, Wegner DM (2010) Causes and consequences of mind perception. Trends Cognit Sci 14(8):383–388. https://doi.org/10.1016/j.tics.2010.05.006
70. Waytz A, Heafner J, Epley N (2014) The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J Exp Soc Psychol 52:113–117. https://doi.org/10.1016/j.jesp.2014.01.005
71. Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM (1832) van Assen MALM (2016) Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. Front Psychol. https://doi.org/10.3389/fpsyg.2016.01832
72. Wullenkord R (2017) Messung und veränderung von einstellungen gegenüber robotern—untersuchung des einflusses von imaginiertem kontakt auf implizite und explizite einstellungen [measurement and change of robot-related attitudes—investigation of the impact of imagined contact on implicit and explicit attitudes]. Ph.d. dissertation, Bielefeld Univ., Bielefeld
73. Wullenkord R, Eyssel F, Nitsch V (2017) The role of self-disclosure in human–robot interaction. In: 26th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 922–927. https://doi.org/10.1109/ROMAN.2017.8172413
74. Złotowski J, Sumioka H, Eyssel F, Nishio S, Bartneck C, Ishiguro H (2018) Model of dual anthropomorphism: the relationship between the media equation effect and implicit anthropomorphism. Int J Soc Robot 10(5):701–714. https://doi.org/10.1007/s12369-018-0476-5

Benedikt Leichtmann is currently a researcher at the Institute of Industrial Engineering and Ergonomics of RWTH Aachen University. After studying Psychology at Ulm University and during a semester abroad at NC State University (2015), he received his Master's degree in Psychology at Ulm University in 2017. From 2016 to 2018, he worked at the Human Factors department of the Bundeswehr University Munich under supervision of Prof. Verena Nitsch, and continued his research at the Institute of Industrial Engineering and Ergonomics of RWTH Aachen University in 2018 working on his doctoral thesis. His research interests include social-psychological aspects in human-robot interaction and the human-centered evaluation of human-robot collaboration in manufacturing.

Verena Nitsch is currently professor and Director of the Institute of Industrial Engineering and Ergonomics at RWTH Aachen University. After studying Applied Psychology at Charles Sturt University in Australia and the University of Central Lancashire in the UK, she completed her Master's degree in Industrial and Organizational Psychology at Manchester Business School. From 2008 to 2013, she worked as a research assistant at the Human Factors Institute at the Bundeswehr University Munich, where she received her Doctorate of Engineering in 2012 and was appointed assistant professor of Cognitive Ergonomics in October 2013. From 2016 to 2018 she was Acting Head of the Human Factors Institute of the Bundeswehr University Munich. Her research interests include human-machine interaction and user-centered technology development.