Social Psychology and Human-Robot Interaction: An Uneasy Marriage

Bahar Irfan
CRNS, Plymouth University, UK
bahar.irfan@plymouth.ac.uk

James Kennedy
CRNS, Plymouth University, UK
james.kennedy@plymouth.ac.uk

Séverin Lemaignan
CRNS, Plymouth University, UK
severin.lemaignan@plymouth.ac.uk

Fotios Papadopoulos
CRNS, Plymouth University, UK
fotios.papadopoulos@plymouth.ac.uk

Emmanuel Senft
CRNS, Plymouth University, UK
emmanuel.senft@plymouth.ac.uk

Tony Belpaeme
ID Lab, Ghent University, Belgium
CRNS, Plymouth University, UK
tony.belpaeme@plymouth.ac.uk

ABSTRACT

The field of Human-Robot Interaction (HRI) lies at the intersection of several disciplines, and is rightfully perceived as a prime interface between engineering and the social sciences. In particular, our field entertains close ties with social and cognitive psychology, and there are many HRI studies which build upon commonly accepted results from psychology to explore the novel relation between humans and machines. Key to this endeavour is the trust we, as a field, put in the methodologies and results from psychology, and it is exactly this trust that is now being questioned across psychology and, by extension, should be questioned in HRI.

The starting point of this paper are a number of failed attempts by the authors to replicate old and established results on social facilitation, which leads us to discuss our arguable over-reliance and over-acceptance of methods and results from psychology. We highlight the recent "replication crisis" in psychology, which directly impacts the HRI community and argue that our field should not shy away from developing its own reference tasks.

KEYWORDS

Social psychology; social robotics; replication crisis; Human-Robot Interaction; research methodology

ACM Reference Format:
Bahar Irfan, James Kennedy, Séverin Lemaignan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2018. Social Psychology and Human-Robot Interaction: An Uneasy Marriage. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3173386.3173389

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03…$15.00
https://doi.org/10.1145/3173386.3173389

1 THE REPLICATION CRISIS IN PSYCHOLOGY AND WHAT IT MEANS FOR HRI

The field of Human-Robot Interaction (HRI), and in particular, the field of *social* HRI benefits from a wide range of scientific input [4, 5]. As a community, we recognise that the technical fields of engineering, control theory and computer science do not provide necessary tools for the scientific investigation of the 'human' and 'interaction' parts of HRI. For this reason, we take inspiration and ground much of our research in established results from the social sciences – primarily social psychology, cognitive psychology, and sociology. As scholars in HRI we find ourselves at the intersection of these many fields, and aim to offer insights to programmers and engineers, as well as psychologists. In this sense, our field embodies the basic idea of cognitive sciences: building bridges across disciplines to gain new insights on complex scientific challenges.

That said, the demographics of the academics working in HRI are skewed towards engineering backgrounds (Table 1); one often becomes a researcher in HRI by first building robots and then looking at how the machines might interact with humans. While some of us do have training in psychology, many do not. This is not an issue per se: as trained scientists and engineers, we can read and interpret the social science literature, and reproduce tasks, protocols, and –perhaps– results.

However, the recent replication crisis in psychology now casts doubt on that premise. Aarts et al. [1], in their seminal study, found that upon attempting to replicate 100 psychology studies, only 39% of the replication studies could subjectively be rated to have replicated the original result. As the results of two thirds of 100 studies could not be properly replicated, whatever the reasons might be (from publication bias, to sociological changes in the population,

Table 1: Academic fields of accepted authors at HRI17, as judged by their affiliation or, if advertised on their personal website, training, $n = 193$ (a single author can be affiliated with multiple fields).

Field	*Eng.*	*Psy.*	*Cog. Sci.*	*Interaction Design*	*Other*
N	145	24	6	17	13
%	70.73	11.71	2.93	8.29	6.34

to small effect sizes), it calls for exerting caution whenever we build upon supposedly established results.

Further research has shown that many scientific studies are difficult or impossible to replicate upon subsequent investigation. According to a 2016 poll of almost 1,600 scientists reported in the journal Nature, over 70% had failed to reproduce at least one other scientist's experiment. More than half had failed to reproduce one of their own experiments [3]. This is problematic for the field of Human-Robot Interaction, as much of what we do either uses research methods similar to those used in other disciplines (and psychology in particular), or relies directly on insights and results handed down from other disciplines.

Because many of us are *consumers* of the psychology literature rather than *producers* or active contributors to the psychology community, we often do not only have insufficient training to correctly interpret psychological studies, but also tend to be less critical and often do not question findings the way we would in our own community. This effect is reinforced by the perceived maturity of different academic fields. Fields such as social or cognitive psychology are very mature, compared to the relative immaturity of Human-Robot Interaction, and studies and insights from psychology are now core material in textbooks, giving the studies and their results further credence.

While experienced researchers in HRI might already be aware of these issues, the influx of new talent requires our field to be vigilant of uncritical reliance on questionable methods and results. To illustrate our point, we present our experience in which we were unsuccessful at reproducing the *social facilitation* effect. Social facilitation, also known as the audience effect, is a supposedly well-established effect where the mere presence of a (silent, passive) external agent influences one's behaviour, often measured through performance on a task. The direction in which the effect works is not specified: depending on the task and the context, performance can be positively or negatively impacted. A large body of literature from psychology reports this effect, and social facilitation has been studied in robotics as well in various forms.

2 A CASE IN POINT: SOCIAL FACILITATION

2.1 Context: Studying the Mere Presence Effect in Social Facilitation

Background and Related Theories. In 1898, Triplett [37] observed that cyclists pedal faster in the presence of rivals than when they are alone. He later studied this effect on children by using a fishing reel that they needed to turn as quickly as possible and found the same effect, although a later analysis of his work by Stroebe [34] showed that there was no significant difference in either of his findings. This effect has later been termed as 'social facilitation' by Allport [2] to describe the increase in response due to the presence of others who are performing the same task. Later the term social facilitation was expanded to cover two types of conditions: 'co-action effects' like Triplett's examples, and 'audience effects', in which only the mere presence of an observer affects the performance of a person performing the task. In order to explain the audience effects, Zajonc [42] proposed the *drive theory*, which states that the audience enhances the exhibition of dominant responses in a person. In the case of a well-mastered task ('simple task'), the performance is facilitated, whereas, for the tasks that are new or require learning ('complex tasks'), the performance is inhibited.

Factors. A meta-analysis by Bond and Titus [8] compared 202 published and 39 unpublished studies on social facilitation. They provide a list of 13 factors that might impact social facilitation (like the participants' age, the number of observers, the role of the observers, the familiarity of the observers, etc.). The meta-study shows that the performance speed (*quantity*) is increased for the simple tasks and the performance accuracy (*quality*) is decreased for the complex tasks. The performance quantity is measured by the latency to respond, time it takes to complete a task and the number of responses per unit time. The performance quality is measured by the number of errors. The analysis also showed that the visibility (presence in the same room as the subject) of the observers has a slightly larger effect than the non-visibility (e.g., one-way mirror [11, 14], use of a video camera [16, 36], a desktop image on a computer screen [15]), although the difference was not statistically significant.

On the other hand, Guerin [17] argues in his review that the majority of studies on social facilitation had observers watching the subject perform a task. These could be confederates, but often they are just the experimenter watching a subject, as they were not seen being busy with other tasks. He also draws attention to ceiling and floor effects of the tasks, and advises that the task should be sufficiently hard so that a reasonable comparison can be made between subjects and conditions.

Tasks. Following Zajonc [42], the literature on social facilitation distinguishes between 'simple tasks' and 'complex tasks'. Examples of simple tasks include cancelling specific letters in a text or multiplication; examples of complex tasks include concept formation, anagrams, digit span, and pursuit rotor tasks (a motor task in which the subject has to track a rotating target using a computer mouse). Tasks such as letter copying and paired associates can be either simple or complex depending on the task structure. McCaffrey et al. [23] also presented significance levels of each of these tasks in the literature. They show that visual perception and construction tasks such as letter or word copying [15, 18, 36] and motor tasks such as physical activities [35] are good tasks in terms of significance as simple tasks. Memory or learning tasks such as paired associates [10, 16, 17] and visuomotor tasks as in the rotary pursuit task [22, 25] have higher significance for social facilitation as complex tasks.

Cheating as a reinforcing factor. Self-presentation theory [7] also suggests conformity to normative behaviours to gain approval of another person. For example, in the case of an embarrassing situation such as cheating, this should prevent the subject from engaging in the cheating behaviour due to social pressure. There might be several factors that affect cheating behaviour, such as the importance of the task, the risk of being caught, the probability of success [39], the belief in free-will [40], the knowledge of peer performance [19], the potential gain of money or grades, the penalty for cheating [26] or conformity to cheating behaviour in peers [13]. In the study by Vohs and Schooler [40], the task consisted of a computer-based mental arithmetic test. The participants were told that there was a "glitch" in the program which shows the correct

answer to the problem, but they could close the answer window by pressing a key after the problem appeared. They were also told that the experimenter would not know whether they pressed the bar, but they should try to solve the problems without looking at the answer. The results revealed that those who were given an essay prior to the test that stated the lack of free-will cheated more frequently than others.

Social Facilitation In Robotics. The audience effect has been studied in HRI by Schermerhorn et al. [32] and Riether et al. [27]. Schermerhorn et al. [32] compared the effect of the robot's presence during easy and difficult arithmetic tasks with alone and robot-presence conditions. A significant two-way interaction between gender and robot was found, because the subjects performed worse during the difficult task when the robot was present. Overall, a marginally significant effect of robot presence was found. Riether et al. [27] on the other hand, compared alone, human-presence, and anthropomorphic robot-presence conditions with four different tasks with easy and complex conditions: anagram solving, numerical distance, finger tapping and a motor reaction task. They observed that in the anagram solving, numerical distance and finger tapping tasks, there were significantly larger performance scores than the alone group for both the robot and human conditions, but there was no significant difference between the robot and the human observer conditions. Authors concluded that this finding suggests that people regard robots as social beings. After the experiment, the subjects were asked to complete a questionnaire in which they gave higher observation impression scores for the robot condition than the human observer, perhaps due to the fact that they thought someone else was watching through the eyes of the robot or due to novelty effects leading to distraction.

Following the findings from social facilitation literature, we decided to explore the mere presence of two robotic platforms (the Softbank Robotics NAO and Pepper robots) through a social facilitation task. We anticipated that there would be a difference between the two platforms due to their size and appearance. While the studies aimed to compare the social facilitation of two different robots, it was important to establish two baselines first: one with no observers, and one with the social facilitation elicited by the presence of a human observer. This would essentially be a first step in validating our methodology and would also serve as replication of the finding from psychology. Assuming the replication study was successful, we would have continued the experiments with a robot as observer and would have compared these results to the earlier obtained baselines.

We ran two separate studies, with a total of three different tasks. Because no effect could be found between the alone condition and the human condition in any of our tasks, we did not actually pursue the studies with robots.

2.2 Social Facilitation: First Attempt

The first study was run between-subjects with two conditions: an alone condition and a human-presence condition. Participants were recruited on a university campus and taken to a room in the campus library for the experiment. The experimenter would take the participant to the room and tell them to follow instructions on the tablet, then the experimenter would leave. In the human observer condition, a second experimenter would already be sitting in the room and would remain there for the duration of the experiment (as per Figure 1).

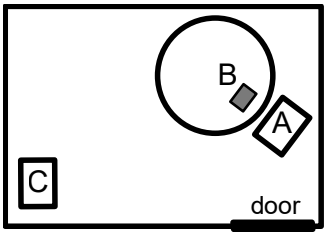

Figure 1: Layout of the room. The participant (*A*) is sitting at a table, with their back to the door. The tasks are performed on a tablet (*B*). When present, the social agent (human observer) is placed at *C*.

Tasks. The literature distinguishes between the effects of mere presence on *simple tasks* and *complex tasks*. We sought to elicit differences in both of these task types. Each participant therefore performed two tasks, followed by a brief questionnaire. Both the tasks and the questionnaire were administered on the tablet. The first task was designed to be a repetitive visuomotor task (the 'shape matching' task); the second one required recollection and comprehension of spoken information (the 'story' task). As such, we examined the effect of social facilitation on both low- and high-cognitive tasks.

The 'shape matching' task is a game where the participants are asked to match a coloured target shape with another one, of the same shape, but of a different colour (Figure 2). The target shape as well as the eight possible responses are random combinations from the sets {red, yellow, green, purple, blue, white} and {square, cross, star, circle}. After the participant touches a shape to select it as an answer, a new random set is shown on screen. This is repeated 200 times. By using the same random seed for all participants, the stimuli sequence was kept identical for all participants.

The task can be repeated for up to 200 rounds of random shapes. After 75 rounds, a button labelled "Give up" appears on screen, giving the participant the option to skip to the second task. The wording of the label was intentionally chosen instead of a more neutral "Stop" or "Continue to next task" to elicit a stronger social response ("Giving up" being more socially costly than simply "continuing to the next task"), thereby increasing the contrast between conditions (self-presentation effect). During this first task, we recorded three metrics: the reaction time for each round, the number of correct and incorrect responses, and the total number of rounds completed. We also asked the participants to give an estimate of how many rounds they thought they had completed, between 0 and 300.

The second task ('story' task) involves listening to a short pre-recorded text (1min 56sec) and answering eight questions about this text. The text[1] details the history of a fictional country named "Brookland" and includes a range of facts: names of places ("[they] sailed to Port Danford"), dates ("Springland was settled in the year

[1]Recording and transcript available on-line, at https://github.com/severin-lemaignan/shapes-matching/tree/master/audio.

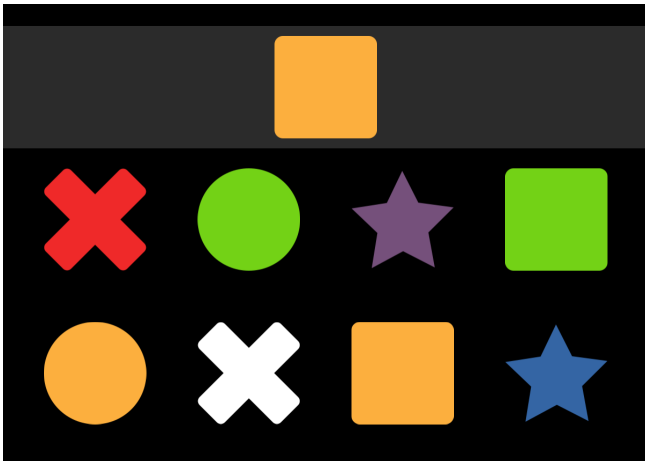

Figure 2: Screenshot of the *shape matching* task. Participants are instructed to tap on the picture matching the target's shape (seen at the top), but with a different colour. In this example, the participant has to tap the green square.

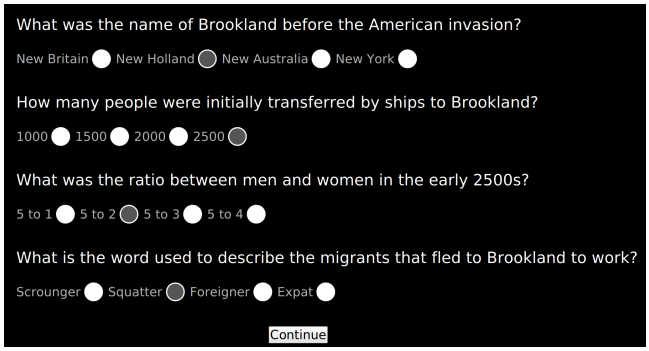

Figure 3: Screenshot of four of the eight questions administered immediately after listening to a short text about a fictional country.

2503"), terminology ("Settlers or 'squatters' began to move deeper into the territories"), and situations ("Women were outnumbered five to one"). The text was based upon the settling of Australia, but with key details and place names changed. This was so that the information would certainly be novel, without sounding implausible. The eight multiple-choice questions are asked immediately after the end of the text. Each question provides a choice of four answers (Figure 3). The score of each participant (number of correct answers) is the performance metric for this task.

Hypotheses. Based on the *drive* theory by Zajonc [42], our hypotheses were the following:

- H1 In the 'shape matching' task, the presence of a social agent would lead to **better performance**: fewer mistakes, faster reaction times.
- H2 In the presence of a social agent, the 'Give up' button would be used less frequently (or later in the game) due to the social pressure (self-presentation theory).
- H3 In the presence of a social agent, participants would report that they completed fewer rounds of shape matching than they actually did, due to social facilitation.
- H4 In the 'story' task, the presence of a social agent would lead to the impaired performance, i.e., participants would recall fewer facts.

Protocol & Data Collection. We recruited 45 participants after exclusion (25 for the alone condition and 20 for the human condition, 16 males, 29 females, balanced across conditions) on campus. The participants' age was M=20.4 (SD=2.5). We ensured that all participants who enrolled were not colour-blind (due to the necessity of seeing colour accurately for the shape matching task) and that they were native English speakers (to prevent comprehension issues due to language in the story task).

Participants were first given information sheets describing the experiment (simply entitled "Learning with a touchscreen", so as not to disclose the role of the mere presence of the observers). They then gave consent to participate, compliant with the university ethics committee rules. Participants were told in writing and verbally that whether or not they decided to withdraw early from the study, they would receive financial compensation of £5 (in the form of a voucher). We made this point explicit to make sure the participants knew that, even if they quit the shape matching game early (i.e., between rounds 75 and 200), they would still receive the full amount.

Results. We did not observe any significant difference between the two conditions concerning the time required to match 75 shapes, average reaction time, number of shape completed, ratio of correct matching, recall performance, or perceived observation (cf. Table 2).

This means that we did not observe any social facilitation effect, and none of the hypotheses are supported.

2.3 Second Attempt

Reflecting on the lack of effect observed in our first attempt, we designed a second experiment to address the possible failures of the first one.

Specifically, we chose (1) to have the human observer closer to the participant (aiming for greater human influence), (2) a stronger moral component (aiming for a greater influence of the human presence), (3) a more difficult task (stronger incentive for behavioural differences – i.e., cheating – between conditions), (4) financial reward dependent on performance (stronger, clearer incentive for behavioural differences between conditions) and finally, (5) regarding the methodology, we decided to move away from primarily using reaction times as metric, so as to avoid any natural performance limit.

Task. Based on these constraints, we designed a new task involving mental arithmetic. Participants were required to calculate the result of a set of non-trivial mental additions. The additions each had exactly three 2-digit numbers to sum, one carry (a digit that is transferred from one column of digits to another), and their results ranged from 100 to 200. Participants had 5 minutes to perform as many additions as possible. Each correct answer would earn them a small financial reward of £0.20 (Figure 4).

Table 2: Results for the shape matching task: time to match 75 shapes, average reaction time, number of shapes completed, ratio of perceived matching, recall performance, and perceived observation. No significance has been observed for any of the metrics (2-tailed independent 2-samples test with equal variance assumption).

Metric	Alone condition $M(SD)$	Human condition $M(SD)$	*p-value*	*t-value*
Time to 75 shapes (s)	117.7 (30.01)	110.63 (17.25)	.349	0.948
Average reaction time (s)	1.70 (0.47)	1.58 (0.26)	.305	1.037
Number of shapes completed	196 (11.5)	198 (7.8)	.522	−0.596
Ratio of correct responses	0.98 (0.02)	0.99 (0.01)	.082	−1.813
Recall performance	4.81 (1.27)	5.11 (1.49)	.473	−0.724
Perceived observation	2.76 (1.27)	2.55 (1.39)	.6	0.528

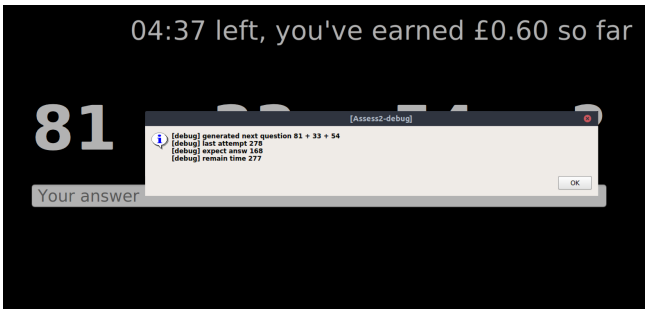

Figure 4: Screenshot of the *sums* task. Participants have 5 minutes to perform as many mental additions of three numbers (with a result between 100 and 200) as possible. Each correct answer earns £0.20. A pop-up dialogue with the answer is shown before each new addition. The participants are instructed that this pop-up is a bug and that they should ignore it and dismiss it.

Critically, following the design of Vohs and Schooler [40], a supposed "glitch" was showing a pop-up dialogue before each addition. This dialogue was designed to look like a spurious debug dialogue and contained the expected answer. The participants were explicitly shown by the experimenter that the correct answer was erroneously displayed in the dialogue. They were instructed to ignore the dialogue and to dismiss it. This 'bug' was explained to the participant as being caused by a new operating system on the laptops used for the test ("Our previous computers did not have this issue"). The bug made it practically easy for participants to cheat: by briefly glimpsing at the debug dialogue before dismissing it, they could immediately know the correct answer, and earn money faster.

The dialogue could be dismissed by pressing 'enter' on the keyboard. 'Enter' was also the key used to move to the next question. As such, a double-press would move to the next question and close the dialogue before it could be seen. Through this mechanism, it was possible to measure how long it took participants to close the dialogue, and infer whether they had cheated.

Hypotheses. The literature suggests that social presence during a complex cognitive task like this one should lead to worse performance [12, 41]. Accordingly, our hypotheses were the following:

- H1 In the presence of a social agent, participants will be more honest (i.e., they will look at the answer on the dialogue pop-up less).
- H2 In the presence of a social agent, participants will complete fewer correct questions.

Protocol & Data Collection. As outlined previously, while our plan was to run four conditions (alone, human presence, NAO presence, Pepper presence), we first ran the two baseline conditions: alone and with a human observer. 15 participants were recruited in the alone condition, 16 participants in the human condition.

The experimental setup was similar to Figure 1 with two differences: when present, the human observer was sitting at the table, facing the participant, and the tablets were replaced with laptops with a keyboard to facilitate the input of the answers. For each participant, we recorded how many additions were attempted, the total gain (i.e., the number of correct answers), and the time to calculate each of the additions. We also asked, upon completion, to what extent they felt like they were being observed during the test (marked on a 5-point Likert scale).

Results. Based on the data (31 participants for a total of 633 additions), the average time to dismiss the debug dialogue was 1185ms and the average time to provide an answer was 9980ms. Based on these values, we conservatively consider cheating as taking more than 0.8 seconds to dismiss the spurious debug dialogue *and* taking less than 5 seconds to calculate the sum *and* providing a correct answer. It results in 147 cheating rounds (23.2% of all rounds).

Looking at these results per condition, we find 77 rounds involving cheating from 316 rounds in the human condition (24.4%) and 70 rounds involving cheating from 317 rounds in the alone condition (22.1%). A 2-samples test for equality of proportions reveals no significant difference; $\chi^2 = 0.463$, $p = .496$. This indicates no support for the presence of a human impacting the tendency to cheat. This result shows that participants do cheat relatively often, however the presence of a human observer does not significantly impact the cheating behaviour of the participants, providing no support for H1.

In terms of performance, participants in the human presence condition gave 28 wrong answers out of 239 rounds with no cheating (11.7% were wrong answers), while participants in the alone condition gave 25 wrong answers out of 247 (10.1%). Using a 2-samples test for equality of proportions, we obtain: $\chi^2 = 0.096$, $p =$

.757 indicating no support that the presence of a human impacts the performance in the test. Again, there is no significant performance difference between the two conditions, providing no support for H2. Therefore, neither of our hypotheses are supported. Due to the absence of any effects between the human and alone conditions, we did not pursue the study with robots.

Participants were asked how observed they felt on a in a Likert scale (1: "Not at all", 5: "Very much"). In the **alone** condition they felt more observed ($M = 2.69$, $SD = 1.35$) than the participants in the human presence condition ($M = 1.75$, $SD = 1.06$). A 2-tailed independent 2-samples test with equal variance assumption shows that there is significant difference: $t(28) = 2.179$, $p = .038$.

3 DISCUSSION

We can only speculate about which factors might explain our failure to observe any effect of social facilitation: the small effect sizes of social facilitation, the setting in which we collected our experimental data, or a bias towards publishing only positive results [31] that might mask how brittle social facilitation effects really are.

Participants also reported the they felt observed: in the alone condition this was $M = 2.69, SD = 1.31$, while in the human condition $M = 1.75, SD = 1.03$. When no experimenter was in the room, they felt *more* observed then when there was an experimenter in the room. This is a very notable result warranting further exploration.

The challenges of observing social interaction. What does this failed attempt at reproducing a "classic" result of social psychology tell us? Beyond possible experimental confounds, our failure at reproducing these results is likely due to the small effect size of social facilitation. In their meta-analysis of studies on social facilitation, Bond and Titus [8] showed that the overall mean effect sizes are low, ranging from 0.03 to 0.36. Uziel [38] reports weighted average effect sizes of less than 0.2. According to Cohen [9], an effect size of 0.2 should be regarded as small, an effect size of 0.5 as medium, and 0.8 as large.

Social facilitation or inhibition, like many other psychological effects, may be affected by a combination of several other factors: the observer effect (also known as the Hawthorne effect [28]), demand characteristics, cultural differences and personality. These effects are potential confounds, and adequately accounting for each of these in the experimental design is problematic.

One likely explanation is that subjects felt observed in both conditions, irrespective of a human observer sitting with them in the room. Just the process of taking part in a study might already exert a large degree of social facilitation, which is not measurably weakened or strengthened by the absence or presence of an observer in the experiment room.

The study of Guerin [18] is relevant in this context: it tried to separate the effect of observer presence from evaluation apprehension. For this a letter copying task was used in four conditions: alone; with a confederate sitting in front of the subject, but facing away; with a confederate at a desk that is behind the subject; and with a confederate sitting behind the subject with no desk in between. Guerin's results showed that there were no significant differences of errors in copying (*quality*) in any conditions, however, alone and front conditions combined were significantly different from the behind and behind-desk conditions combined in terms of task performance (*quantity*).

Furthermore, he used self-reports for determining the level of pressure the subjects felt. Subjects in the alone condition were asked to imagine how they would feel if there was a person in the room. The results showed that the subjects in the alone condition felt *more* disturbed and evaluated than those of the other three conditions, which concurs with the results we found. However, he noted that self-reports in social facilitation research may be affected by demand characteristics and self-presentation. As a result of the study, he was unable to separate evaluation apprehension from the mere presence effect on task performance.

It is likely that the subjects in our study felt observed by taking part in a study. Even though the true intent of the study was not revealed until the debriefing, subjects felt observed whatever the condition and this might have impacted their behaviour. This is know as the Hawthorne effect. However, the Hawthorne effect itself is a subject of discussion as there are studies that challenge its existence. Jones [21] studied the original experiment data [28], and found that there is slight or no evidence of a Hawthorne effect. McCambridge et al. [24] reviewed over 19 studies that investigate the Hawthorne effect, and argued that the term is used to describe a broad range of effects in the literature rather than the core definition which refers to the change in subjects' behaviour due to conformity to perceived norms or researcher expectations. Hence, they could not confirm whether the effect exists.

Weak methods in older psychology literature. Beyond the caution that must be observed when studying one specific psychological effect, a broader range of methodological issues with older research in psychology might explain why some results in psychology are incorrectly believed to be reliable.

For instance, the Bond and Titus [8] meta-analysis of research on social facilitation claims to have exhaustively examined every publication prior to the publication of the meta-analysis itself (in 1983). As a matter of fact, the oldest study that they reference dates from 1898, and 35 out of the 241 were published prior to 1965. As such, social facilitation is a good example of an old, classical psychological effect. It however also hints at the fact that its characterisation might have relied on weak research methodologies by today's standards. In that regard, Bond and Titus raise interesting points: only 100 out of the 241 studies state that the experimenter was in a different room in the alone condition (and in 96 studies, we know the experimenter was in the room). This would be seen today as a serious confound. Similarly, Bond and Titus report that 72.3% of the total participants were undergrad students, pointing to a possible demographic bias.

Biases in scientific publishing: the 'file drawer' problem. Coined in 1979 by Rosenthal [30], the *file drawer problem* refers to the bias introduced into the scientific literature by mainly publishing positive results, and rarely negative or non-confirmatory results. As a consequence, an effect could be reported and believed reliable, simply for the lack of literature showing the contrary. Rosenthal proposes to account for this problem by reporting in meta-analysis the 'fail-safe N' measure: N is the number of null effects that would be required to make the original result non-significant. Rosenthal

considers an effect resistant to the 'file drawer problem' of unreported null effects if the fail-safe N is above $5k + 10$, with k the number of reported effects.

Bond and Titus [8] report the fail-safe N for some of the effects of social facilitation. For instance, their meta-analysis show that the performance quantity of participants for complex tasks reliably decreases in presence of an observer (even thought the effect size is small). 54 effects are reported, and they note that the fail-safe N value is 160: 160 is clearly smaller than $5 \times 54 + 10 = 280$ and as such, this result could well be subject to the problem of unreported null effect. The fact that social presence inhibits the performance in complex tasks is not a robust result in the face of the bias towards publishing only positive results.

A weighted calculation of the fail-safe number has been proposed [29] that addresses some of the concerns with Rosenthal's proposal, and while not systematically reported in the literature, this metric is a valuable tool for HRI researchers when assessing how robust a result in psychology is.

4 CONCLUSION

While we have built this paper around social facilitation and our failed attempt at replicating this well-established effect, the observations we make above are broadly applicable to Human-Robot Interaction. Our failure to replicate a result from social psychology which has stood for 120 years [37] should form a cautionary tale. The limited reproducibility of results in psychology seems to be endemic [1] and while the reasons for the lack of reproducibility are many and diverse, there is a genuine concern that the field of HRI is also affected. We are, however, not suggesting that HRI should not build upon psychology anymore. Quite the contrary. Our field has strong ties with psychology, and our work is grounded in various theoretical and methodological frameworks. If anything, we encourage the community to keep on building new links with neighbouring academic fields, and social psychology should be a preferred partner in this effort.

However, we need to be frank: results from social psychology, experimental methodologies and reporting methods which were considered as commonly accepted or even gold standards until recently, are losing their special status. Instead we would like to offer the following suggestions to the HRI field:

Replicate and reproduce. When replicating a social psychology effect with robots, it is necessary to first reproduce the effect with people. Methods change, times and mores change, and negative results often go unreported. A social psychology effect which is touted in textbooks might not be that easy to replicate. With psychology at the centre of the recent replication controversy, many results which seem established should be approached with the necessary skepticism.

Null-results are interesting. The field of HRI most likely also suffers from publication bias and the file drawer effect: many studies go unreported because the results are inconclusive, negative or because they do not support an agenda. If results are negative or insignificant, the field needs to know. This helps us focus our resources better: if an experiment returned negative results and we know about it, then it can help us avoid setting up a similar experiment. It also helps us with quantifying bold claims. As results come in that are inconclusive or unsupportive of those claims, they tend to go unpublished or do not get the same amount of airtime and attention as confirmatory results. This culture should change.

Avoid questionable research practices. A number of questionable research practices (QRPs) have been identified in social psychology [20, 33]. While we have not collected data on the presence of QRPs in HRI, we need to be aware of the QRPs identified in psychology. Examples include (from [20]) selective reporting of data, or only reporting data which support a particular story; collecting data until the results are significant; p-value rounding, i.e., rounding p-values down to .05 to suggest statistical significance (a particular problem of null-hypothesis testing); failing to report all conditions; or selectively reporting studies that "worked".

Register your study. In clinical studies, it is customary to register the study protocol before beginning data collection (see for example clinicaltrials.gov). Perhaps a similar practice should be established for HRI. Among the many benefits, the registration of trials before running include the reduction of publication bias, the efficient allocation of research resources, and full engagement with ethical obligations of the research community.

Avoid the Hawthorne effect. The set-up of most HRI studies often reveals to the subjects that they are being observed: lab-based studies always implicitly signal to subjects that their behaviour will be monitored. Even moving into a naturalistic environment might not alleviate this problem, as ethics procedures insist that subjects are briefed before a study and that their explicit consent is sought before they can engage in the experiment. As such, subjects in HRI experiments might always experience the Hawthorne effect: their behaviour changes because they are aware of being observed. The only way forward here is to either not inform subjects prior to the study (which is unethical) or work with a distractor task. However, the latter is particularly difficult to implement in HRI.

Come up with HRI reference tasks. While there is merit in attempting to reproduce effects from social psychology with robots instead of people, it might be worth identifying new effects and tasks relevant to Human-Robot Interaction and its applications. Times change and as robots become more ubiquitous, our response to robots is likely to evolve rapidly. We need to look at the relation and interaction between people and robots through new lenses, and the old (often very old) views from social psychology are perhaps no longer applicable or appropriate. It may be noted that our implementation of the methodology did not perfectly match one from psychology. The task used in the second attempt was the same as one from psychology, however, we did not deploy the essay writing portion of the original [40] so as not to introduce a confound. Finding an appropriate methodology to replicate in the context of HRI was a challenge in itself, further reinforcing the need for our own reference tasks.

As a community, HRI should learn from its own mistakes (see Baxter et al. [6] for good advice) and from the mistakes of others. We are a young community, with a steady influx of young talent, and we often look towards established fields for guidance. But when exactly these established fields start to question their own practices

and results, we should too. The conclusion of the Science study on reproducibility in psychology [1] offers the following message:

> Following this intensive effort to reproduce a sample of published psychological findings, how many of the effects can we confirm are true? Zero. And, how many of the effects can we confirm are false? Zero. Is this a limitation of the project design? No. It is the reality of doing science, even if it is not appreciated in daily practice.

Importantly, this is the reality of doing science *in general*, not only *social* science. We must not blind ourselves: our methods and protocols in HRI do not shelter us from the exact same problems experienced in other fields. Future researchers may well write the same kind of article about our field when they revisit today's literature on Human-Robot Interaction.

ACKNOWLEDGMENTS

This work has been partially supported by the EU H2020 L2TOR project (grant 688014), the EU FP7 DREAM project (grant 611391), the EU H2020 Marie Sklodowska-Curie Actions Innovative Training Networks project APRIL (grant 674868), and the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227). All authors have contributed equally to the experimental design, execution, data analyses and writing.

REFERENCES

[1] A.A. Aarts, C.J. Anderson, J. Anderson, M.A.L.M. van Assen, P.R. Attridge, et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015). https://doi.org/10.1126/science.aac4716

[2] F.H. Allport. 1924. *Social Psychology.* Houghton Mifflin Company, Chapter Response to social stimulation in the group, 260–291.

[3] M. Baker. 2016. 1500 scientists lift the lid on reproducibility. *Nature News* 533, 7604 (2016), 452. https://doi.org/10.1038/533452a

[4] C. Bartneck. 2011. The End of the Beginning: A Reflection on the First Five Years of the HRI Conference. *Scientometrics* 86, 2 (2011), 487–504.

[5] C. Bartneck. 2017. Reviewers' scores do not predict impact - Bibliometric Analysis of the Proceedings of the Human-Robot Interaction Conference. *Scientometrics* 110, 1 (2017), 179–194. https://doi.org/10.1007/s11192-016-2176-y

[6] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme. 2016. From characterising three years of HRI to methodology and reporting recommendations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 391–398.

[7] C.F. Bond. 1982. Social Facilitation: A self-presentational view. *Journal of Personality and Social Psychology* 42, 6 (1982), 1042–1050.

[8] C.F. Bond and L.J. Titus. 1983. Social facilitation: a meta-analysis of 241 studies. *Psychological bulletin* 94, 2 (1983), 265–292. https://doi.org/10.1037/0033-2909.94.2.265

[9] J. Cohen. 1977. *Statistical power analysis for the behavioral sciences*. Academic Press.

[10] N.B. Cottrell, R.H. Rittle, and D.L. Wack. 1967. The presence of an audience and list type (competitional or noncompetitional) as joint determinants of performance in paired associates learning. *Journal of Personality* 35 (1967), 425–434.

[11] W.D. Criddle. [n. d.]. The physical presence of other individuals as factor in social facilitation. *Psychonomic Science* 22, 4 ([n. d.]), 229–230.

[12] J.M. Feinberg and J.R. Aiello. 2010. The Effect of Challenge and Threat Appraisals Under Evaluative Presence. *Journal of Applied Social Psychology* 40, 8 (2010), 2071–2104.

[13] T.R. Fosgaard, L.G. Hansen, and M. Piovesan. 2013. Separating Will from Grace: An experiment on conformity and awareness in cheating. *Journal of Economic Behavior and Organization* 93 (2013), 279–284. https://doi.org/10.1016/j.jebo.2013.03.027

[14] V.J. Ganzer. 1968. Effects of audience presence and test anxiety on learning and retention in a serial learning situation. *Journal of Personality and Social Psychology* 8, 2 (Pt. 1) (1968), 194–199.

[15] W.L. Gardner and M.L. Knowles. 2008. Love Makes You Real: Favorite Television Characters Are Perceived as "Real" in a Social Facilitation Paradigm. *Social Cognition* 26, 2 (2008), 156–168. https://doi.org/10.1521/soco.2008.26.2.156

[16] R.G. Geen. 1973. Effects of being observed on short- and long-term recall. *Journal of Experimental Psychology* 100 (1973), 395–398.

[17] B. Guerin. 1983. Social Facilitation and social monitoring: a test of three models. *British Journal of Social Psychology* 22 (1983), 203–214. https://doi.org/10.1111/j.2044-8309.1983.tb00585.x

[18] B. Guerin. 1989. Reducing evaluation effects in mere presence. *The Journal of Social Psychology* 129, 2 (1989), 183–190.

[19] J.P. Hill and R.A. Kochendorfer. 1969. Knowledge of peer success and risk of detection as determinants of cheating. *Developmental Psychology* 1, 3 (1969), 231–238. https://doi.org/10.1037/h0027330

[20] L.K. John, G. Loewenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* 23, 5 (2012), 524–532.

[21] S.R.G. Jones. 1992. Was There a Hawthorne Effect? *Amer. J. Sociology* 98, 3 (1992), 451–468.

[22] J.P. Lombardo and J.F. Catalano. 1975. The effect of failure and the nature of the audience on performance of a complex motor task. *Journal of Motor Behavior* 7 (1975), 29–35.

[23] R.J. McCaffrey, J.M. Fisher, B.A. Gold, and J.K. Lynch. 1996. Presence of third parties during neuropsychological evaluations: Who is evaluating whom? *The Clinical Neuropsychologist* 10, 4 (1996), 435–449. https://doi.org/10.1080/13854049608406704

[24] J. McCambridge, J. Witton, and D.R. Elbourne. 2014. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology* 67, 3 (2014), 267–277.

[25] F.G. Miller, M.E Hurkman, J.B. Robinson, and R.A. Feinberg. 1979. Status and evaluation potential in the social facilitation and impairment of task performance. *Personality and Social Psychology Bulletin* 5 (1979), 381–385.

[26] D.S. Nagin and G. Pogarsky. 2003. An Experimental Investigation of Deterrence: Cheating, Self-Serving Bias, and Impulsivity. *Criminology* 41, 1 (2003), 167–194. https://doi.org/10.1111/j.1745-9125.2003.tb00985.x

[27] N. Riether, F. Hegel, B. Wrede, and G. Horstmann. 2012. Social facilitation with social robots?. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12*. 41–47. https://doi.org/10.1145/2157689.2157697

[28] F.J. Roethlisberger, W.J. Dickson, H.A. Wright, and Western Electric Company. 1939. *Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago.* Harvard University Press.

[29] M.S. Rosenberg. 2005. The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution* 59, 2 (2005), 464–468.

[30] R. Rosenthal. 1979. The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86, 3 (1979), 638–641.

[31] H.R. Rothstein, A.J. Sutton, and M. Borenstein. 2006. *Publication bias in meta-analysis: Prevention, assessment and adjustments.* John Wiley & Sons.

[32] P. Schermerhorn, M. Scheutz, and C.R. Crowell. 2008. Robot Social Presence and Gender: Do Females View Robots Differently than Males? *ACM/IEEE International Conference on Human-Robot Interaction* (2008), 263–270. https://doi.org/10.1145/1349822.1349857

[33] J.P. Simmons, L.D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

[34] W. Stroebe. 2012. The truth about Triplett (1898), but nobody seems to care. *Perspectives on Psychological Science* 7, 1 (2012), 54–57.

[35] M.J. Strube, M.E. Miles, and W.H. Finch. 1981. The social facilitation of a simple task: Field tests of alternative explanations. *Personality and Social Psychology Bulletin* 7 (1981), 701–707.

[36] D.J. Terry and M. Kearnes. 1993. Effects of an audience on the task performance of subjects with high and low self-esteem. *Personality and Individual Differences* 15, 2 (1993), 137–145.

[37] N. Triplett. 1898. The dynamogenic factors in pacemaking and competition. *American Journal of Psychology* 9, 4 (1898), 507–533.

[38] L. Uziel. 2007. Individual differences in the social facilitation effect: A review and meta-analysis. *Journal of Research in Personality* 41, 3 (2007), 579–601. https://doi.org/10.1016/j.jrp.2006.06.008

[39] F.T. Vitro and L.A. Schoer. 1972. The effects of probability of test success, test importance, and risk of detection on the incidence of cheating. *Journal of School Psychology* 10, 3 (1972), 269–277. https://doi.org/10.1016/0022-4405(72)90062-3

[40] K.D. Vohs and W. Schooler. 2008. The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating. *Psychological Science* 19, 1 (2008), 49–54. https://doi.org/10.1111/j.1467-9280.2008.02045.x

[41] I. Wechsung, P. Ehrenbrink, R. Schleicher, and S. Möller. 2014. Investigating the Social Facilitation Effect in Human-Robot Interaction. In *Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice*, J. Mariani, S. Rosset, M. Garnier-Rizet, and L. Devillers (Eds.). Springer, New York, 167–177.

[42] R.B. Zajonc. 1965. Social facilitation. *Science* 149, 3681 (1965), 269–274.