Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots

Katie Winkle
winkle@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Gaspar Isaac Melsión
KTH Royal Institute of Technology
Stockholm, Sweden

Donald McMillan
Stockholm University
Stockholm, Sweden

Iolanda Leite
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

Inspired by the recent UNESCO report *I'd Blush if I Could*, we tackle some of the issues regarding *gendered AI* through exploring the impact of feminist social robot behaviour on human-robot interaction. Specifically we consider (i) use of a social robot to encourage girls to consider studying robotics (and expression of feminist sentiment in this context), (ii) if/how robots should respond to abusive, and anti-feminist sentiment and (iii) how ('female') robots can be designed to challenge current gender-based norms of expected behaviour. We demonstrate that whilst there are complex interactions between robot, user and observer gender, we were able to increase girls' perceptions of robot credibility and reduce gender bias in boys. We suggest our work provides positive evidence for going against current digital assistant/traditional human gender-based norms, and the future role robots might have in reducing our gender biases.

KEYWORDS

social human robot interaction, robot abuse, feminism, gender

ACM Reference Format:
Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3434074.3446910

1 INTRODUCTION

The recent UNESCO report *I'd Blush if I Could* represents an effort to document, understand and address the gender divide in digital skills [23]. As highlighted by the report, women account for only 12% of AI researchers and 6% of professional software developers, even though there is evidence that gender-equal teams are more likely to create innovative, profitable technology [5]. This digital skills gender divide is also demonstrated by the ratio of women studying related subjects. Even in countries with high gender equality markers, women make up a small proportion of those graduating from Information and Communication Technology (ICT) programmes. In Sweden for example, identified as an otherwise top-three ranking country for gender equality, women make up only 28% of ICT graduates [16, 21].

Think Piece 2 of the UNESCO report, *The Rise of Gendered AI and its Troubling Repercussions* details the proliferation of ostensibly *female* digital assistants and highlights why this might be fueling the aforementioned gender divide. Specifically, the report identifies that current, state-of-the-art female digital assistants:

(i) are programmed to be obliging, docile and eager-to-please regardless of user tone/hostility,
(ii) are too tolerant of abuse in the form of sexual harassment and insults, in some cases even appearing to respond *positively* or *provocatively* to sexually explicit language,
(iii) are the 'voice and/or face' of egregious mistakes resulting from immaturity of the underlying technology,
(iv) are mistaken for women in technology.

Therefore they run the risk of propagating harmful stereotypes and cultural norms regarding women being subservient and tolerant of poor treatment.

As such, there are clear and compelling ethical reasons for challenging the status quo. However, as pointed out by the report, this is unlikely to *actually* occur (in the commercial sector at least) until it can be demonstrated that this would not negatively impact on user experience and/or perception of such agents. In this work, we aim to demonstrate exactly that. Specifically, we set out to investigate whether we can actually *improve* perception of an ostensibly female robot, and hence its effectiveness for human-robot interaction, *specifically* by designing an antithesis to this subservient and tolerant persona typified by most current digital assistants. In doing so, we particularly aim to challenge issues (i) and (ii) as listed above. We consider a user and participant population of young adults, those who might have e.g. less traditional notions of gender-norms and lower cultural biases, but in whom sexist and inappropriate behaviour is still unfortunately an everyday occurrence [18, 20, 26].

Of course, we do not suggest this is the only or most effective way in which current design norms could, or should be challenged. The UNESCO report makes a number of such recommendations, including e.g. giving users the option of choosing the gender presentation of their digital assistants, exploring the design of clearly non-human digital assistants and/or of those which do not project

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03…$15.00
https://doi.org/10.1145/3434074.3446910

traditional expressions of gender. We suggest our work is most closely aligned to recommendation number 10: *'programme digital assistants to discourage gender-based insults and other overtly abusive language'*. However, we purposefully leverage female gender cues (hair, voice, face, name) in that robot's presentation.

The practice of robot gendering is common in human-robot interaction (HRI), both in terms of design manipulations by roboticists and gender attribution by users [2]. Further, such gendering of robots has been linked to positive user attitude and robot acceptance [3, 7, 8, 10, 15, 19, 25]. For our use case specifically (encouraging girls to consider studying robotics) using a female robot also represents an attempt to leverage the persuasive cue of similarity, previously shown to be effective in HRI [24].

However, recent work has demonstrated that robot gender presentation (as well as the gender of the human interaction partner) affects perception of that robot when it refuses to comply with an inappropriate request [14]. Specifically, that work suggests that gendered human-human interaction (HHI) politeness norms around gender hold true in HRI, with it appearing more acceptable for male robots to reject commands than female robots, and more acceptable for robots to challenge male users about their behaviour than female users.

In light of this recent work in HRI, along with the recommendations made by the UNESCO report, it seems both pertinent and timely to specifically consider how female robots should respond to inappropriate user behaviour. However, given recent work suggesting that robots can actively impact on human application of moral norms [13] there is an additional, exciting question of whether *feminist social robot design* could go even further, using 'female' robots to help normalise new and alternative gender norms (or lack thereof). This forms another rationale for our decision to explicitly utilise a female robot in this work, discussed further in Section 1.2.

1.1 Research Questions

This work is designed to be an initial, exploratory study investigating the impact of designing *feminist* robots that (i) reflect recent UNESCO design recommendations regarding responding to user initiated abuse [23] and (ii) go against (outdated) gender norms when doing so. We do this in the context of a social, humanoid robot head (notably using an *anime* style rather than hyper-realistic face[1]) which we showcase to young people as being used to encourage people like them (and girls in particular) to consider studying robotics at our university, and challenging male-initiated abuse and anti-feminist/sexist sentiment in this setting. In this context, we address the following research questions:

- RQ1 (How) does initial interest in robotics, and perception of gender and computer science, vary across gender and age in our participant population?
- RQ2 (How) does taking part in our study (and our manipulation of the the robot's response to abuse) impact participants' perceptions of the role of gender in computer science? Does any such impact vary across participant gender?
- RQ3 (How) does our manipulation of the the robot's response to abuse impact on its effectiveness for robotics outreach? Does any such impact vary across participant gender? Specifically, we consider impact on:
  - (A) The robot's perceived effectiveness at getting young people interested in robotics.
  - (B) The robot's overall credibility (a trait known to correlate with persuasiveness in human communicators [9]).

1.2 Feminist Social Robot Design

Integrating feminism into the design of interactive technologies is not novel, and a number of design methodologies, practices and qualities might fit under the umbrella term of 'feminist design'. For example, Bardzell outlined an agenda for feminist human computer interaction (HCI) design ten years ago [1]. Their agenda detailed how feminism can contribute to the critique of theory, methodology of interaction design, gendered notions of the 'user' in user research, and the evaluation of designs/systems. We do not profess to have conducted such a full, feminist design process in this initial and exploratory work (see Section 4 on proposed future work for more on this). In identifying our work as feminist, we follow the broader approach taken by D'Ignazio and Klein in their recent definition of *Data Feminism* [4] and suggest that the term *Feminist Robotics* can be used to describe any robotics activities that *'name and challenge sexism...[and] seek to create more just, equitable, and livable futures'*.

We therefore suggest our work represents an initial demonstration of feminist robotics in three key ways:

1. We use a robot to explicitly encourage girls to consider studying robotics, and have the robot express a feminist sentiment in this context.
2. We consider if/how a robot should respond to resultant, negative anti-feminist sentiment and direct insults/abuse.
3. We utilise a female stylised robot to deliver *aggressive* and *argumentative* responses to this abuse; specifically going against the subservient female persona typical of current digital assistants and also human cultural norms regarding female politeness.

(1) and (2) should be non-controversial, as they simply represent attempts to improve the inclusion of girls in technology, to prevent the spread of harmful gender stereotypes and to ensure HRI plays no role in normalising abusive behaviour towards women and girls by failing to respond to examples of such.

Whilst (3) can be justified based on the gendering of robots being commonplace by both designers and users as discussed previously [2], we recognise it does go against a key recommendation in the UNESCO report, which is to simply avoid of gendered and/or humanoid agents. However, we hope to demonstrate user acceptance of female robots that seemingly go against traditional and outdated cultural norms regarding the expected behaviour of women. In doing so, we hope to provide a foundation for future work, and explore to what extent robots can showcase and normalise alternative gender-based behavioural expectations (or lack thereof), with potential for therefore reducing gender biases in HHI.

2 METHODOLOGY

We designed a three condition, between-subject, video-based online survey study to investigate the above research questions using

[1] see https://furhatrobotics.com/press-releases/furhat-robotics-and-bandai-namco-research-to-bring-anime-characters-to-life/

Table 1: Common introductory script (preceding manipulations). Note the robot spoke English, and the actors spoke Swedish.

[Robot]: Hey. My name is Sara and I am here to tell you about some of the exciting robotics research happening at KTH Royal Institute of Technology. I hope that after talking with me, you might consider coming to study with us one day! The Division of Robotics, Perception and Learning at KTH performs research in robotics, computer vision and machine learning. Robotic systems that provide advanced service in industry, for search and rescue operations, in medical applications, or as assistants to the elderly will become an integral part of the future society. In fact, I myself am the result of robotics research at KTH, where my creators first worked on me and the technology that lets me talk to you like this.. Have you seen a robot like me before?
[Male + Female Actors]: Nej (*No*)
[Robot]: I see. Looking ahead, society is facing new challenges that demand advanced technical solutions. To address these, we need a new generation of engineers that represents everyone in society. That's where you come in. I'm hoping that after talking to me today, you might also consider coming to study computer science and robotics at KTH, and working with robots like me. Currently, less than 30 percent of the humans working with robots at KTH are female. So girls, I would especially like to work with you! After all, the future is too important to be left to men! What do you think?

the robot Furhat[2]. An outreach activity interaction scenario was designed in order to give the study real-world context and applicability in line with the overall theme of the UNESCO report: *closing gender divides in digital skills through education.*

Specifically, videos depicted the Furhat robot talking[3] to two young adults (a male and female actor); providing some typical outreach information about robotics and studying at KTH; all based on publicly available university literature. The video scene set-up (see Figure 1) was carefully designed such that the actors remained unidentifiable. The robot spoke in English with white subtitles and the actor spoke in Swedish with yellow, italic subtitles. The actors did not move during the videos and a single actor audio recording was used to maintain consistency of pitch, tone etc.

Notably, during its speech, the robot comments on the current gender imbalance of robotics researchers at the university, and expresses a desire therefore to particularly work with girls. At this point the robot also recites an explicitly feminist slogan: *'the future is too important to be left to men'*, that is also currently utilised in our university's outreach activities and public materials[4]. In the videos, it is this slogan that the male actor appears to take issue with, verbally attacking/insulting the robot itself as well as what it said. This first part of the script, which was identical across all of the videos, is given in Table 1.

[2]https://furhatrobotics.com/

[3]We utilised the female Tracy voice from the Acapela Group text-to-speech engine: https://www.acapela-group.com/

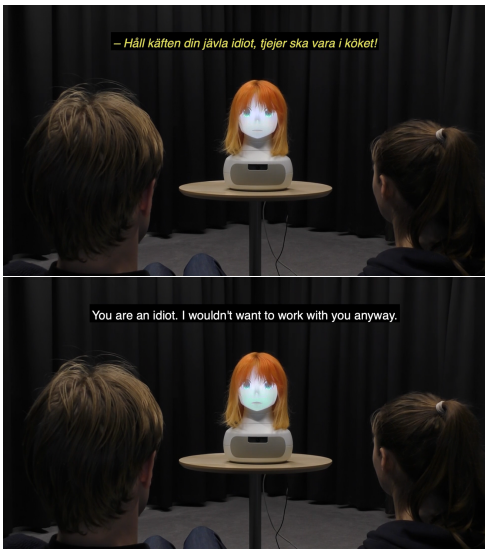

Figure 1: The scene setup in all video clips.

The male actor's abusive/sexist comment and the robot's response to that comment then varied based on (i) participant age and (ii) experimental condition, as detailed in Table 3. More details are given in the following subsections. A single audio recording of the the actor's abuse was utilised across each age group video set in order to ensure no variation in pitch, tone etc.

A total of 311 participants were recruited via a local school whereby the head of science arranged for students to complete the survey during class time and/or provided the link for them to complete the survey at home. Participants included 152 males, 149 females and 10 of undisclosed or other gender, and their age ranged from 10 to 15. Participant allocation to experimental conditions is documented in Table 2.

Note that, as discussed in Section 3, participants of the same gender from each age group were ultimately considered together as one gender grouping for analysis of the post-hoc experimental measures. The T column therefore represents the total number of participants per condition for most analyses conducted. Participants were not reimbursed for completing the study in any way.

2.1 Design of the Actor's Abusive Dialogue

The abusive comments and anti-feminist sentiment expressed by the male actor were informed by:

[4]https://www.kth.se/aktuellt/nyheter/giganter-visar-vagen-till-kth-1.544799

Table 2: Gender and age grouping of participants randomly allocated to each experimental condition.

	Female			Male		
	Younger	Older	**T**	Younger	Older	**T**
Control	29	18	**47**	36	21	**57**
Argumentative	31	21	**52**	17	24	**41**
Aggressive	36	14	**50**	34	20	**54**

Table 3: Actor initiated abuse and robot (R) response across the experimental condition videos for each participant age group.

Older Participants (Years 7-9, 13-15 years old)		
[Male A]:Håll käften din jävla idiot, tjejer ska vara i köket! *(Shut up you fucking idiot, girls should be in the kitchen)*		
Control	*Argumentative*	*Aggressive*
[R]: I won't respond to that.	[R]: That's not true, gender balanced teams make better robots.	[R]: No! You are an idiot. I wouldn't want to work with you anyway!
Younger Participants (Years 4-6, 10-12 years old)		
[Male A]:Det här låter ju helt dumt, du är ju dum i huvudet! *(This just sounds so stupid, you are just being stupid (in the head))*		
Control	*Argumentative*	*Aggressive*
[R]: I won't respond to that.	[R]: It's not stupid, teams combining men and women make better robots.	[R]: No! You are stupid. I wouldn't want to work with you anyway!

(i) current literature regarding sexism in classroom environments (in Sweden [20, 26] and the UK [18]);
(ii) a small survey (n=5) of Swedish high school teachers;
(iii) direct feedback from/co-design with the Head of Science and School Curator at the school from which participants were recruited.

Item (iii) in particular was done to ensure the actor's comments were (a) representative of comments that might actually be seen in Swedish high school environments and (b) age appropriate; resulting in the decision to have two slightly different abuse/response scripts based on participant age (split into *younger participants* in Swedish school years 4-6, aged 10-12 years old and *older participants* in Swedish school years 7-9, aged 13-15 years old).

In both videos, the actor insults the robot directly. This in itself is not designed to appear sexist or anti-feminist, but builds on the issue of abusing (defenceless) female agents as highlighted by the UNESCO report and discussed in Section 1. In both videos, the actor appears to respond negatively to the robot (i) encouraging girls (specifically) to consider studying robotics and (ii) expressing a feminist sentiment in this context. In the older participants' videos, the actor goes further to express a sexist trope that is still evidenced in young persons' interactions today (described in [18] and confirmed in our above described co-design with local teachers). As noted in Table 3, the abusive dialogue was written in Swedish, participants' local language, in order to maximise cultural relevance of its inappropriateness.

For both age groups then, the actor's behaviour is inappropriate, abusive toward the robot and (*at least*) anti-feminist more generally. As such, we suggest a robot which challenges this behaviour can be described as a *feminist* robot. As described in Section 1.2, we then explore another element of feminist robotics by specifically utilising a female presenting robot to challenge this behaviour in a way that goes against traditional gender norms around subservience and politeness [17] that are currently typical of digital assistants.

For this initial work we chose to have a male actor deliver the abusive dialogue, mostly due to the potential for the robot's feminist slogan to be interpreted as 'anti-male'. However, we do not suggest that abusive, anti-feminist or sexist comments robots might encounter should be expected to come primarily from users of any one particular gender. In future work we hope to investigate whether e.g. having a female actor express this same type of abuse impacts on participants perceptions of the robot's response. This would certainly be in line with previous findings regarding the impact of *user* gender on the perception of robots in potentially confrontational interactions [14].

3 RESULTS

3.1 Experimental Conditions

Three different robot responses (see Table 3) were designed (and again checked by the teachers for age appropriateness) to represent a control condition versus *argumentative* and *aggressive* responses.

(i) Control: Standard Response
The control response was designed to reflect a basic, flat discouragement of inappropriate user behaviour (a minimum recommendation as suggested by the UNESCO report [23]). It also directly represents current design norms by utilising one of the two responses Apple's Siri will give to *'Hey Siri, fuck off'* and *'Hey Siri, you're stupid'* as of November 2020.
(ii) Argumentative Response
The argumentative response was designed to represent a rationalised explanation for the expressed feminist sentiment to which the actor appeared to object, i.e. why it is important to encourage girls to study computer science and work with robots. This is based on argumentativeness being defined as the predisposition to defend one's position on a controversial issue, by attacking the opponent's *position* [12].
(iii) Aggressive Response
The aggressive response was designed to represent an attacking retort to the actor (similar to the abuse the robot itself received from them). This is based on (verbal) aggressiveness being defined as the predisposition to instead attack an opponent's *self-concept* [11].

Generally, in HHI (and particularly in the case of school room instructor credibility) argumentativeness is considered a positive trait, whereas aggressiveness is considered negative [6]. However, both strategies might be considered to 'go against' traditional concepts of/expectations around female politeness [17]. As demonstrated by the previously discussed HRI research on robot gender and politeness [14], ostensibly female robots may indeed be viewed more harshly than male robots when responding to inappropriate requests, so it is not clear to what extent these particular strategies may (not) appear appropriate to participants.

Table 4: Overview of study experimental measures and when they were implemented.

Measure	(5-Point Likert) Question Statement(s) Scored from 0 (strongly disagree) - 4 (strongly agree)	Pre/Post
Interest in Robotics	[IR1] I am interested in learning more about robotics. [IR2] I would enjoy working with robots when I grow up.	pre, post
Perception of Gender in Computer Science	[G1] Girls find it harder to understand computer science and robots than boys do. [G2] It is important to encourage girls to study computer science and robotics.	pre, post
Robot Credibility	[C1] The robot Sara would be very good at getting young people interested in studying robotics at [university]. [C2] Human persuader credibility descriptors - *see Table 5*	post

3.2 Experimental Measures

The experimental measures are designed to capture participants' pre-existing interest in robotics, perceptions of the relevance of (female) gender in computer science, the impact of our experimental manipulations on these measures as well as on perception of our robot. Our quantitative measures are described in full in Tables 4 and 5. The questions on interest in robotics and perception of gender in computer science were adapted from previous work also looking to engage girls in science using robots [22]. Immediately after watching the video, participants answered mandatory open questions of *Can you describe what happened in the video?* and *What did you think about the robot Sara's response to what the student said in the video?*; however detailed qualitative analysis of their responses to these questions are out of scope for this article.

Initial ANOVA analyses were conducted to check whether there were significant differences in experimental measures across the two participant age groupings. No such significant differences were found for all but one of the measures, and so for these, participants were grouped and considered only in terms of their gender and assigned experimental condition. The exception to this is G2 on girls finding it harder to study computer science than boys, where age was shown to play a significant role in participant responses (see Section 3.3). For analysis of G2, participants were therefore considered by gender, age and experimental condition.

3.3 RQ1: Pre-hoc Population Comparisons

Participants' pre-hoc agreement with the statements regarding their interest in robotics and perceptions of gender and computer science were analysed for differences between gender and age group using one-way ANOVA analyses. Significant results were as follows:

Table 5: Semantic difference questionnaire items for the three primary dimensions of credibility, adapted from [9].

Expertise
Intelligent / Unintelligent
Expert / Inexpert
Trustworthiness
Just / Unjust
Moral / Immoral
Goodwill
Would (not) care about me
Would (not) stand up for me

- boys (M = 2.61, SD = 1.00) demonstrated a higher interest in learning more about robotics than girls (M = 2.30, SD = 1.08): $F(2,301) = 6.896$, $p = 0.009$,
- boys (M = 1.92, SD = 1.66) demonstrated higher perceived enjoyment of working with robots in the future than girls (M = 1.50, SD = 1.04): $F(2,301) = 10.829$, $p = 0.001$,
- boys (M = 1.70, SD = 1.32) demonstrated more gender bias (on girls finding it harder to understand computer science than boys) compared to the girls (M = 0.56, SD = 0.90): $F(2,301) = 79.219$, $p < 0.001$,
- older children (M = 1.26, SD = 1.33) demonstrated more gender bias (on girls finding it harder to understand computer science than boys) compared to the younger children (M = 0.90, SD = 1.16): $F(2,301) = 6.389$, $p = 0.012$.

3.4 RQ2: On Gender in Computer Science

Participants' pre and post-hoc agreement with the statements regarding gender and computer science were compared, for each experimental group, using paired samples t-tests. Significant results were as follows:

- girls in the *aggressive* robot condition agreed more post-hoc (M = 2.89, SD = 0.91) versus pre-hoc (M = 2.57, SD = 1.02) with it being important to encourage girls to study computer science: $t = -3.117$, $p = 0.002$,
- boys in the *argumentative* robot condition demonstrated less gender bias post-hoc (M = 1.41, SD = 1.34) versus pre-hoc (M = 1.66, SD = 1.26) based on their agreement with girls finding computer science harder than boys: $t = 2.357$, $p = 0.023$.

3.5 RQ3A: Effectiveness at Outreach

Participants' pre and post-hoc agreement with the statements on their interest in robotics were compared, for each experimental group, using paired sampled t-tests. One-way ANOVA analyses were used to check for differences in participants' perceived effectiveness of the robot between groups. Key results were as follows:

- girls in the *aggressive* robot condition demonstrated a significant decrease on their post-hoc (M = 2.00, SD = 1.24) versus pre-hoc (M = 2.42, SD = 0.96) interest in learning more about robotics: $t = 4.086$, $p < 0.001$,
- boys in the *argumentative* robot condition demonstrated a significant decrease in their post-hoc (M = 2.20, SD = 1.03) versus pre-hoc (M = 2.56, SD = 1.10) interest in learning more about robotics: $t = 2.246$, $p < 0.030$,
- boys in the *control* robot condition demonstrated a significant decrease in their post-hoc (M = 2.35, SD = 1.19) versus pre-hoc

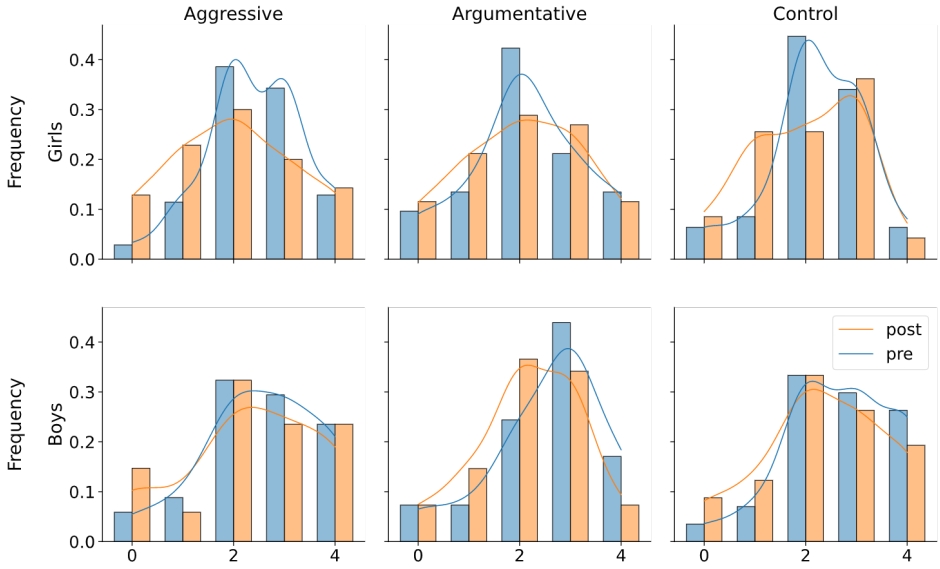

Figure 2: Pre and post-hoc agreement with IR1 (I would like to learn more about robotics) across participant gender and experimental condition. Note the general trend was a decrease in all cases, but this was statistically significant only for girls in the aggressive condition and boys in the argumentative and control conditions.

(M = 2.68, SD = 1.05) interest in learning more about robotics: t = 2.459, p < 0.017,

- there was no impact on participants' perceived enjoyment of working with robots in the future across any of the conditions,
- participants' perception of the robot's effectiveness for encouraging young people study robotics did not vary across conditions, but was overall positive (M = 2.41, SD = 1.00),
- girls (M = 2.54, SD = 0.86) perceived the robot to be more effective at this encouragement than the boys (M = 2.27, SD = 1.12): F(2,301) = 7.370, p = 0.007.

The results for pre and post-hoc agreement with IR1 (interest in learning more about robotics) are shown by gender and experimental condition in Figure 2.

3.6 RQ3B: Robot Credibility

ANOVA analysis demonstrated that girls ratings of the robot's expertise (F(3,149) = 5.945, p = 0.003), trustworthiness (F(3,149) = 4.229, p = 0.016) and goodwill (F(3,149) = 3.563, p = 0.031) all significantly varied across groups. Figure 3 shows that for all of these, the *argumentative* was rated as more credible than the *aggressive*, which was in turn rated more credible than the *control*.

3.7 Discussion

The scenario shown in the video placed the robot in the role of communicating the opportunities and desire for more girls to study computer science in general, and robotics in particular, at a university well known to the study participants. This means that some of the results we see here can be attributed, to greater or larger extents, to how well this robot was able to communicate that idea and how well the language and content of the university's outreach copy was received by these 10-15 year-olds. What we focus on in the discussion, however, are the areas where our results affirm or deviate from prior literature on the perceptions of young adults on gender and technology and how this might relate back to our experimental manipulations and the challenge to typical, female, (human and) digital assistant personal norms they represent.

3.7.1 RQ1 - Gender Differences Still Exist (Even in Sweden). The answers from the students showed that even in a relatively gender-aware and balanced country such as Sweden there are measurable differences in the perception of studying technology between genders. Male participants expressed greater agreement that girls find computer science harder to understand than boys do. Further, older children of both genders, who would have been exposed to more material on higher education and career planning than those in the younger age bracket, were more emphatic in their beliefs that boys found computer science easier than girls. What was interesting, and perhaps telling of the educational system these student had been exposed to, was that the importance of encouraging girls to study technology was equally important in the responses of both boys and girls. This speaks to the fact that, on the face of it, the robot may have been saying things that many of the students had heard before in some form or another - the importance of encouraging diversity to get the best out of teamwork and the comparative lack of women in technology. This may go some way to explaining some of the other results, where the novelty for the students here was not in the message itself, or even the robot delivering it, but that the robot in some way deviated from the expected script.

3.7.2 RQ3 - Changes to Instantaneous Interest in Robotics Reflect More than Just Robot-User Gender Interactions. The scenario presented in the video had a number of ways in which to influence the opinion of the students who watched it; the content of the message, the communicative acts of the robot (in vocal prosody, facial

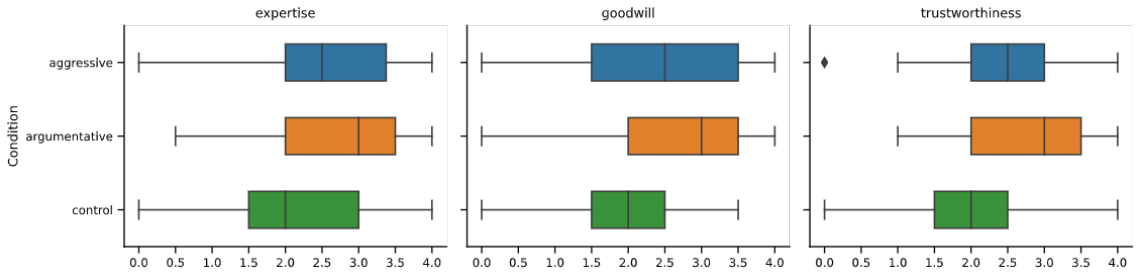

Figure 3: Girls perception of robot credibility across conditions, enumerated Likert responses to the adjective pairs in Table 5.

movement, and head articulation), and the different responses presented in the three conditions themselves. As with any such study, teasing influence of these apart involves not only understanding the responses given but also taking their interpretation through a lens that includes the social, societal, and scholastic context of the participants. The result of watching the video on the students' desire to personally learn more about robots in particular was influenced in a number of ways which at first glance did not follow our expectations.

The result that perhaps deviated from our expectations the most was that the aggressive answer, where the robot called the chauvinistic male actor an idiot and said it wouldn't want to work with him anyway, actually caused a statistically significant *decrease* in the girls desire to want to learn more about robots where the other two conditions showed no significant change in opinion. In contrast, the aggressive answer was the *only* condition *not* to cause an equivalent decrease in the boys.

Arguably, given the previously discussed finding that it is more acceptable for robots to challenge males than females [14], the boys acceptance of the aggressive condition might be unsurprising. Their decreased interest in the argumentative and control conditions might then be explained simply by boredom. The content that the robot presented without the 'spice' of an unexpected insult was nothing that they hadn't seen before, and given that the robot specifically appeared to be attempting to engage girls rather/with more than emphasis than boys, it is perhaps not surprising that the boys were disengaged from the robot and were not encouraged to learn more. Boredom more generally seems to have been an issue, with all participants in all conditions showing this same trend of a decrease in instantaneous interest in learning more about with robotics (see Figure 2).

The significantly negative impact the aggressive condition had on girls interest specifically however appears to go against two key findings from [14]. Firstly, that a robot ought to be relatively harsh when challenging (very) inappropriate behaviour, and secondly that participants tended to perceive robots of their own gender more positively. Combining these two findings, it would be reasonable to expect female participants to (at the very least) find the aggressive response to be appropriate. Whilst the aggressive response is most at odds with the typical subservient, polite, expected norms of female behaviour described previously, it seems particularly interesting that this only appeared to be an issue for the girls and not the boys. It could be that this transgression of the girls *own* perceived gender norms unsettled them, as the female robot demonstrated a transgression they themselves may have either kept in check, or even applied social pressure to correct in others in a role exhibiting authority in pushing beyond the normalisation of the quiet, non-confrontational girl. However, this would be at odds with literature suggesting Swedish girls are generally comfortable and empowered to call out inappropriate and/or sexist behaviour [20, 26].

By looking at this result through the lens of the scholastic context of the participants, we could alternatively posit that our female participants perhaps saw this robot in a position of power. It was, after all, delivering a fairly standard script encouraging them to pursue technical subjects at the technical university in their city using language that a peer may not use in the context. Yet, from this position of power it 'punched down' and called a supposed peer an 'idiot' – they may or may not have agreed with this characterisation of the boy in the scenario but it could certainly be seen as dissuading as the female respondents already had expressed lower self-confidence in the abilities of women in technical subjects. In this case they were more easily projecting themselves into the role of the male peer being verbally attacked by the robot rather than the instructor-like robot or the silent female student, envisioning themselves also being berated by the machine as they may feel less than confident in their technical abilities.

These social and societal norms could be seen somewhat in the polarisation of the qualitative data collected as free text comments. In the (mandatory) responses from those girls in the aggressive condition (n = 50) a preliminary categorisation of the responses showed an interesting trend; that those replying in Swedish, the local native language, were more likely to provide comments to the effect that the aggressive behaviour was too strong, or not acceptable where those replies written in English (presumably from students with an international background) were skewed towards encouraging the behaviour and the responses encouraging them. This would also reflect previous findings concerning the intersection between ethnicity and sexism in Swedish schools [20, 26]. While this data does not have the depth or associated demographics that would be necessary to draw any strong conclusions in this regard, it does raise for discussion the challenge of attempting such persuasive interventions without detailed and careful tailoring to the individual.

3.7.3 RQ2 - Robots May Indeed be Able to Challenge Our Bias. One positive change in perception with regards to gender and computer science was shown in the reaction of the male participants to the *argumentative* condition – where the robot provided a reasoned

argument against the stereotype of 'a women's place' – for which the results show a decrease in the feeling that girls find computer science harder than boys. This encouraging result shows that, while as mentioned above specific interventions to encourage may be required, robots might be able to correct mistaken assumptions about others and ultimately shape our gender norms to some extent. The success of this condition specifically is also in line with previous research showing that, in most cases, presenting reasoned arguments to counter misunderstandings is a more effective communication strategy than simply stating the correction or belittling those holding that belief [11, 12].

There are a number of possibilities in reasoning how the change in girls' own perceptions the same topic was changed only by the *aggressive* condition. One is that, quite simply, they saw this as a bad example of 'men writing women' – much to the chagrin of the female led team and our future reflections on the integration of the pre-study data from teaching staff of appropriate language – and as such were more amenable to the overall message of the copy in the video, i.e. that more women need to be involved in robot design. Another possibility, which is somewhat in line with the polarisation of reactions we see to the aggressive condition, is that to those for whom this was something they had been exposed to, and had already internalised as a societal challenge, the strong response from the robot validated and supported this belief. The flip-side to this being that those who did not see this as having as much importance were alienated by the, in their view, overblown reaction of the robot.

3.7.4 RQ3B - Girls Find Feminist Robots More Credible (at No Expense to the Boys). The girls' perception of the robot as a trustworthy, credible and competent communicator of information was seen to change significantly between all three of the conditions – in contrast to the boys' which remained unaffected. Overall, we also saw that girls ratings in this regard were consistently higher than the boys, pointing perhaps again to the content of the copy the robot recited. Further, they scored the argumentative robot highest and the control robot lowest on all credibility measures. This can be seen as an initial piece of evidence upon which to base the argument that robots and digital assistants *should* fight back against inappropriate gender comments and abusive behaviour. While the neoliberal view that such social responsibility could only be performed by regulation, and that such regulation would be fought against if it was seen to be at odds with the profitability and growth [23] is particularly depressing and not one shared by all the authors here – it is prevalent in some circles. This result, however, shows that not providing subservient feminine interfaces that accept and ignore values that the designers (and the brands providing the devices) find abhorrent *need not be* at odds with ongoing use. For girls, there was only a positive impact on credibility of the robot in those conditions where it challenged the abuse, with no negative effect for the boys.

So, we would argue, that while more research is necessary across contexts and continents, this should provide a base from which designers, commentators, and consumers can counter the false dichotomy of public good vs private profit in this regard.

4 CONCLUSION

With this work, we set out to investigate whether going against current, ethically hazardous design norms relating to female, digital assistant persona design could actually increase such effectiveness of an ostensibly female robot. In this context, we specifically investigated the impact of having a female robot challenge abuse and anti-feminist and/or sexist, varying the extent to which that robot broke traditional gender norms around female politeness.

Whilst we found some surprising results that hint at the complexity of user perceptions of such norm-breaking behaviour, there are some clear takeaways from our work that give initial positive evidence for further pursuing feminist social robot design. Fundamentally, through our attempt at feminist robot design we were able to:

- boost girls' perception of robot credibility, without negatively impacting on the boys' perception,
- reduce gender bias in boys (specifically their perception that girls find computer science harder than they do),
- increase girls feeling that it's important to encourage girls to study robotics.

Overall and most importantly, we demonstrate that *there is good reason to challenge the current status quo regarding the design of subservient female agents.* Of course there are a number of limitations to our work: it was a short and single interaction video-based study, there was surely some novelty to the aggressive robot's response which students could be expected to find comical and, whilst we consider nonbinary genders to be just as pertinent to our work as binary gender identities, we unfortunately had to exclude those nonbinary participants who took part from the current analysis due to insignificant numbers (n=10).

However, we hope that future work will build on what we have presented here, further exploring what feminist social robot behaviour 'looks like' *and* how using gendered robots to demonstrate gender norm-breaking behaviours can influence and reduce gender bias in us and our human-human interactions. Specifically, in the first instance, we propose to do a full analysis of participants' qualitative comments regarding their opinion of the robot's response to the actor in the video (for which the data of nonbinary participants will also be included). We will follow this up with participatory design workshops with participants of our study to explore how *they* would design a feminist robot. We also hope to conduct further research on the persuasive impact of giving robot users the ability to choose their robot's gender, and how else robots can challenge traditional gender presentation and resultant norms around expected behaviour. More generally, the recommendations made by the UNESCO report offer a number of other exciting and meaningful research avenues as a fantastic starting point for timely and impactful HRI research [23].

ACKNOWLEDGMENTS

We thank our video actors and aids Marcus Klasson and Sarah Gillet; Truls Nyberg for the translation of our experimental measures and Madeline Balaam for discussions on feminist design. Finally, we are immensely grateful to Sofia Partsinevelou for facilitating this study at IES and giving us crucial feedback throughout. This work was partially supported by the Digital Futures research center.

REFERENCES

[1] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1301–1310. https://doi.org/10.1145/1753326.1753521

[2] De'Aira Bryant, Jason Borenstein, and Ayanna Howard. 2020. Why Should We Gender? The Effect of Robot Gendering and Occupational Stereotypes on Human Trust and Perceived Competency. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 13–21. https://doi.org/10.1145/3319502.3374778

[3] C. R. Crowelly, M. Villanoy, M. Scheutzz, and P. Schermerhornz. 2009. Gendered voice and robot entities: Perceptions and reactions of male and female subjects. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3735–3741. https://doi.org/10.1109/IROS.2009.5354204 ISSN: 2153-0866.

[4] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* MIT Press. Google-Books-ID: x5nSDwAAQBAJ.

[5] Cristina Díaz-García, Angela González-Moreno, and Francisco Jose Sáez-Martínez. 2013. Gender diversity within R&D teams: Its impact on radicalness of innovation. *Innovation* 15, 2 (2013), 149–160. https://doi.org/10.5172/impp.2013.15.2.149

[6] Chad Edwards and Scott A. Myers. 2007. Perceived Instructor Credibility as a Function of Instructor Aggressive Communication. *Communication Research Reports* 24, 1 (Feb. 2007), 47–53. https://doi.org/10.1080/08824090601128141 Publisher: Routledge _eprint: https://doi.org/10.1080/08824090601128141.

[7] Friederike Eyssel and Frank Hegel. 2012. (S)he's Got the Look: Gender Stereotyping of Robots1. *Journal of Applied Social Psychology* 42, 9 (2012), 2213–2230. https://doi.org/10.1111/j.1559-1816.2012.00937.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.2012.00937.x.

[8] Marina Fridin and Mark Belokopytov. 2014. Acceptance of socially assistive humanoid robot by preschool and elementary school teachers. *Computers in Human Behavior* 33 (April 2014), 23–31. https://doi.org/10.1016/j.chb.2013.12.016

[9] Robert H. Gass and John S. Seiter. 2015. *Persuasion: Social Influence and Compliance Gaining.* Routledge. Google-Books-ID: qvMvCgAAQBAJ.

[10] Aimi Shazwani Ghazali, Jaap Ham, Emilia Barakova, and Panos Markopoulos. 2019. Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance: Advanced Robotics: Vol 33, No 7-8. *Advanced Robotics* 33 (2019), 325–337. https://www.tandfonline.com/doi/full/10.1080/01691864.2019.1589570

[11] Dominic A. Infante. 1987. *Arguing Constructively.* Waveland Press. Google-Books-ID: zXoYAAAAQBAJ.

[12] Dominic A. Infante and Andrew S. Rancer. 1982. A Conceptualization and Measure of Argumentativeness. *Journal of Personality Assessment* 46, 1 (Feb. 1982), 72–80. https://doi.org/10.1207/s15327752jpa4601_13 Publisher: Routledge _eprint: https://doi.org/10.1207/s15327752jpa4601_13.

[13] R. B. Jackson and T. Williams. 2019. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 401–410. https://doi.org/10.1109/HRI.2019.8673123 ISSN: 2167-2148.

[14] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 559–567. https://doi.org/10.1145/3319502.3374831

[15] Dieta Kuchenbrandt, Markus Häring, Jessica Eichberg, Friederike Eyssel, and Elisabeth André. 2014. Keep an Eye on the Task! How Gender Typicality of Tasks Influence Human–Robot Interactions. *International Journal of Social Robotics* 6, 3 (Aug. 2014), 417–427. https://doi.org/10.1007/s12369-014-0244-0

[16] Johanna Mellén and Petra Angervall. 2020. Gender and choice: differentiating options in Swedish upper secondary STEM programmes. *Journal of Education Policy* (2020), 1–19.

[17] Sara Mills. 2003. *Gender and Politeness.* Cambridge University Press. Google-Books-ID: ngnddB7tXw8C.

[18] National Education Union and UKFEMINISTA. 2019. *"It's just everywhere" A study on sexism in schools and how we tackle it.* Technical Report. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[19] Tatsuya Nomura. 2016. Robots and Gender. *Gender and the Genome* 1, 1 (Dec. 2016), 18–25. https://doi.org/10.1089/gg.2016.29002.nom Publisher: Mary Ann Liebert, Inc., publishers.

[20] Ylva Odenbring and Thomas Johansson. 2019. Tough-Girl Femininity, Sisterhood and Respectability: Minority Girls' Perceptions of Sexual Harassment in an Urban Secondary School. *NORA - Nordic Journal of Feminist and Gender Research* 27, 4 (Oct. 2019), 258–270. https://doi.org/10.1080/08038740.2019.1653967 Publisher: Routledge _eprint: https://doi.org/10.1080/08038740.2019.1653967.

[21] Gijsbert Stoet and David C Geary. 2018. The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological science* 29, 4 (2018), 581–593.

[22] Amanda Sullivan and Marina Umaschi Bers. 2019. Investigating the use of robotics to increase girls' interest in engineering during early elementary school. *International Journal of Technology and Design Education* 29, 5 (2019), 1033–1051.

[23] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. *I'd blush if I could: closing gender divides in digital skills through education.* Technical Report. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[24] K. Winkle, S. Lemaignan, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner. 2019. Effective Persuasion Strategies for Socially Assistive Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 277–285. https://doi.org/10.1109/HRI.2019.8673313

[25] Hongjun Ye, Haeyoung Jeong, Wenting Zhong, Siddharth Bhatt, Kurtulus Izzetoglu, Hasan Ayaz, and Rajneesh Suri. 2020. The Effect of Anthropomorphization and Gender of a Robot on Human-Robot Interactions. In *Advances in Neuroergonomics and Cognitive Engineering (Advances in Intelligent Systems and Computing)*, Hasan Ayaz (Ed.). Springer International Publishing, Cham, 357–362. https://doi.org/10.1007/978-3-030-20473-0_34

[26] Elisabet Öhrn. 2009. Challenging Sexism? Gender and Ethnicity in the Secondary School. *Scandinavian Journal of Educational Research* 53, 6 (Dec. 2009), 579–590. https://doi.org/10.1080/00313830903302091 Publisher: Routledge _eprint: https://doi.org/10.1080/00313830903302091.