Face to Face with a Sexist Robot

Investigating how women react to sexist robot behaviors

Diljot Garcha[1], Denise Geiskkovitch[2], Raquel Thiessen[1], Susan Prentice[3], Kerstin Fischer[4], James Young[1]

[1]Computer Science, University of Manitoba, Winnipeg, Canada, {garchads, thiess83}@myumanitoba.ca, young@cs.umanitoba.ca
[2]Computing and Software, McMaster University, Hamilton, Canada, geiskkod@mcmaster.ca
[3]Sociology, University of Manitoba, Winnipeg, Canada, susan.prentice@umanitoba.ca
[4]Design and Communication, University of Southern Denmark, Denmark, kerstin@sdu.dk

ABSTRACT

Social robots are often created with gender in mind, for example by giving them a designed gender identity or including elements of gender in their behaviors. However, even if unintentional, such social robot designs may have strong gender biases, stereotypes or even sexist ideas embedded into them. Between people, we know that exposure to even mild or veiled sexism can have negative impacts on women. However, we do not yet know how such behaviors will be received when they come from a robot. If a robot only offers to help women (and not men) lift objects for example, thus suggesting that women are weaker than men, will women see it as sexist, or just dismiss it as a machine error? In this paper we engage with this question by studying how women respond to a robot that demonstrates a range of sexist behaviors. Our results indicate that not only do women have negative reactions to sexist behaviors from a robot, but that the male-typical work tasks common to robots (i.e., factory work, using machinery, and lifting) are enough for stereotype activation and for women to exhibit signs of stress. Particularly given the male dominated demographic of computer science and engineering and the emerging understanding of algorithmic bias in machine learning and AI, our work highlights the potential for negative impacts on women who interact with social robots.

1 Introduction

Social robots are being developed to serve as personal assistants at workplaces, companions and assistants in homes, kiosks in banks and shopping malls, and to provide emotional support in clinics and centers. These robots are designed to behave as social actors, following tendencies to attribute social characteristics to machines (Media Equation, Reeves & Nass, 1996), using human-like language and interaction techniques to work naturally and comfortably with people; a kiosk robot may be equipped with humanoid features and designed to use human-like speech, gaze, and gestures to facilitate interaction. Social robot interaction design must further consider the surrounding social structures and context of interaction (Young et al., 2008) as, just as between people, what is deemed effective or appropriate from a robot changes largely based on factors such as social hierarchies or physical place. One such integral component of this is gender, where the person's gender identity, the robot's designed (or perceived) gender identity, gendered aspects of the interaction situation (e.g., male or female-typical roles), all interrelate to fundamentally shape the resulting interaction (Y. Wang & Young, 2014). Just as between people, matters surrounding gender will impact human-robot interaction. However, as a robot is not human, any designed "pseudo" gender is only analogous to gender in people; we do not yet know entirely how gender will relate to interaction with a social robot. For example, if a robot makes a sexist statement, would the statement be seen as sexist (given that it is coming from a machine) or laughed off as a technical error? In this paper we investigate how prominent gender considerations and issues may impact human-robot interaction. We present the results from a study where we create a gendered situation and varying types of gender-charged interactions and analyze the results of human-robot interaction from a range of gendered perspectives. Specifically, we explore the questions: how will women interact with a robot in gendered contexts, and how does a robot's gendered interactions (which may be sexist) impact this interaction?

Issues surrounding gender are prominent for social robotics, given that Computer Science and Engineering, the primary fields developing robots, remain heavily male-biased (S. Wang & Bunt, 2017).

For the time being we can expect most robots and their behaviors to be created predominantly by men, and thus have a tendency toward male-typical or male-focused ideas, applications, use contexts, and behavior characteristics (Y. Wang & Young, 2014). For example, we informally note the apparent tendency for successful robot technology to *support* male-typical work with men still squarely in the equation as operators (e.g., construction, farming, manufacturing), while many emerging prototypes aim to *replace* female-typical labour (e.g., "nurse" bots, domestic chores). This follows a plethora of historical examples of how male-dominated technology development can lead to biased results that exclude or negatively impact the female population (Adam, 1998; Perez, 2019; Schiebinger, 2008).

Robot designs and behaviors themselves can exhibit and reinforce stereotypes, potentially being sexist. It is common for design to leverage simple stereotypes (e.g., that children like animals) in attempts at effective interaction, but one could also imagine – without any malevolent intent – teams creating robots that by-default consult female family members regarding domestic cooking and cleaning duties, or suggest that fathers consult with a female family member regarding childcare. Ambivalent sexism theory (Glick & Fiske, 1997, 2011) highlights that even seemingly positive or flattering sentiments can have roots in more harmful sexist ideas. For example, a robot assuming that women are more nurturing than men (relating to the idea that women are better at domestic roles), or one should hold doors for women (women need protection by men), also evokes potentially harmful gender stereotypes (cf. Steele 1998). A well-meaning designer could conceivably create a robot that kindly offers to help women with something mechanical (where it may not offer such help to a man) based on similar assumptions. Data-driven approaches can also lead to sexist behaviors (Howard & Borenstein, 2018), where machine learning may reproduce all-too-accurate reflections of existing inequalities, or even create them in an attempt to generalize (Baker & Potts, 2013; Kay et al., 2015). For example, a robot learning from occupation statistics may assume that doctors and pilots it meets are male, or that women are better suited for caregiver jobs.

Thus, it is not only possible, but likely, that robots will both be created with sexist behaviors, and that interaction will be commonly embedded into male-typical use contexts. This creates several problems for human-robot interaction. First, the robot and behavior may reinforce harmful stereotypes such as that women are not technically competent, or men cannot care well for children, with far-reaching consequences (Winkle et al., 2021, 2022). Such biases may hinder interaction comfort, impacting the quality of interaction, both for sexist behaviors, and for women interacting with robots in more male-typical workplaces and contexts. Further, this kind of exposure to sexism can cause short-term reduction in performance on cognitive tasks (Dardenne et al., 2007; Schmader & Johns, 2003), and can work to entrench stereotypes, for example, by making women feel unwelcome or less interested in robotics (Cheryan et al., 2009, Winkle et al., 2021, 2022). With the male-dominated technical workforce, these issues are more likely to be sexist toward women than men.

Given the infancy of social robots, the discussion is predominantly academic, with very little data available about how people react to sexist or gender-charged robot behaviors. In this paper, we investigate if indeed people would respond to a robot's sexist behavior as if it were from a person, or if instead they would perhaps discount it as a machine error. On the one hand, given a robot's autonomy and physical embodiment (and thus a kind of agency, Young et al., 2010) we may expect people to respond to gender constructs and interactions from a robot similar to how they would for a person; this may be amplified given that robot rudeness or mistakes may increase some aspects of anthropomorphism (e.g., attribution of mental state, Short et al., 2010). On the other hand, gender relations and sexism between people are deeply rooted in a long, complex human history involving, among other things, power dynamics and oppression. We do not yet know how social robots will fit into this rich context, or how existing woman-man dynamics, for example, map to similar interactions between a woman and a robot.

We conducted a study that orchestrated gendered interactions between women and a social robot and present an analysis of how these women interacted with and reacted to our robot. That is, we designed a human-robot interaction scenario (drawing from Dardenne, Dumont, & Bollier, 2007), where we recruited 38 female participants to interact with a sexist robot and investigated the impact on participants and their interaction with the robot. Following previous work outlined above, we measured short-term impact of exposure to sexism on cognitive task performance and mood, expecting to see similar negative impacts on both as observed with sexism from a person (Dardenne et al., 2007). We measured impact on anthropomorphism of the robot, to see if prior work on the impacts of impolite robot behavior on attribution of human-like characteristics is reflected (Short et al., 2010). Further, we analyze videos of the interactions for signs of participant stress and disaffiliation toward the robot throughout, both during a male gendered scenario as well as during the sexist behaviors.

Our results indicate that women do respond negatively to male-gendered robot use scenarios of robots, and do perceive sexist robot behaviors as sexist, reacting negatively. However, we did not find a difference in ability on a cognitive task as expected based on prior work. These results reflect on the nuances of the overarching question of how sexist robot behaviors robot's behavior may be received by people and impact them, highlighting the importance of better understanding sexism in robot design. These results provide early insight into human-robot interaction as robots continue to be developed for use throughout society.

2 Background

The field of Science, Technology, and Society studies (commonly referred to as STS) has a rich history of investigating relationships between technology and society, highlighting both how societal norms and concepts influence technology, and how technology itself in turn influences society (Hackett et al., 2007). The intersection of gender and technology has received particular attention, with a wealth of research highlighting the impact of male-dominated technology development on women, society, and the resulting technological artifacts themselves (see, e.g., Adam, 1998; Perez, 2019; Schiebinger, 2008). A key result is that gender is inextricably integrated into technological development and use. Gender, including societal norms, historical and societal context, and individuals' gender identities – and how all of these interact – must be considered carefully to fully understand how any technology will be used and adopted by individuals and society. What is yet unclear is how a robotic social actor will fit into this broad gender context.

2.1 Ambivalent Sexism Theory

While we do not provide detailed treatment of gender from a theoretical or sociological standpoint, we briefly address Ambivalent Sexism Theory (Glick & Fiske, 1997, 2011) as it has particular relevance to human-robot interaction and our research. Work in this area highlights how even subtle, perhaps seemingly well-intentioned sexist behaviors can strongly impact women, both psychologically and in task performance.

Ambivalence refers to having simultaneous opposing thoughts or ideas, and ambivalent sexism is the notion that sexism has simultaneously both a *hostile* and a *benevolent* component. Hostile sexism, perhaps most familiar to readers, is overtly negative, sexist stereotypes relating to gender. A person believing (and perhaps acting on beliefs) that men are better than women at mathematics is an example of hostile sexism as it blatantly judges women negatively based solely on gender or sex. In contrast to hostile sexism, benevolent sexism refers to gender stereotypes which may appear positive or favorable on the surface (especially to the person holding the stereotype) but are in fact harmful and have roots in damaging gender stereotypes[1]. For example, many traditional ideas regarding supporting and celebrating women as caretakers and as being emotionally superior, such as those based in chivalry, often posit that women need physical protection and help (from men), founded in the prejudicial idea that women are weaker and less resourceful than the men who will protect them.

The effects of hostile sexism are broadly studied; for example, research shows that the presence of negative stereotypes can reduce women's math performance (Schmader & Johns, 2003). We may

[1] The authors note that the term "benevolent" is problematic, in that it implies good intention and underlying positivity, although such behaviours often do not have positive intent nor be based on doing actual good. Perhaps the term "honeyed sexism" or "veiled sexism" would better reflect the concept. We continue to use "benevolent" to match the standard in the literature.

reasonably assume that a robot exhibiting related opinions or behaviors will be offensive and received negatively, although no research has yet investigated this (e.g., perhaps people would not react to the robot as seriously). However, benevolent sexism is much more insidious: a series of studies (Dardenne et al., 2007) demonstrated how women did not necessarily identify benevolent sexism *as* sexism, and yet exposure to it had a negative impact on their wellbeing and short term cognitive performance. That is, during a mock job interview scenario, women did not identify a "benevolent sexist" behavior as sexism per se but performed even more poorly on cognitive tests after benevolent sexism than the hostile sexism (which was easily identified).

The reasons behind why benevolent sexism would impact women in this way – worse than hostile sexism – are complex and not fully resolved. Exposure to a stereotype (called stereotype "activation"), even when not explicitly identified as such, can lead to stereotype-conforming and supporting behaviors (Bargh et al., 1996; Gupta et al., 2008). However, the impact can be reduced when the stereotype is overt (as in the hostile case), as people can cognitively consider and counteract the stereotype (Gupta et al., 2008). Similarly, psychological reactance theory (Steindl et al., 2015) notes how people react negatively to threats to their liberties or freedoms with increased motivation to act to protect themselves: thus we may expect additional effort (reactance, e.g., rooted in anger) to counteract the stereotype given hostile sexism, but less reactance for the benevolent case where the threat is less obvious (Dardenne et al., 2007).

There are many reasons why benevolent sexism may be particularly harmful. Dardenne et al. (2007) note that, even if women do not identify benevolent sexism as sexism, it can still make them feel uncomfortable, which in turn can result in an altered mental and motivational state. Further, such subtle or benevolent sexism may be ambiguous, perhaps causing women to exert mental energy focusing on it, struggling to tell whether the behavior was sexist or not, or to consider why they are feeling uncomfortable (Basford et al., 2014; Benokraitis, 1997). We must also consider stereotype threat, where (in our case) in a stereotyped scenario a woman may worry about a stereotype about women, causing anxiety and stress (Spencer et al., 2016). In all, this helps explain why women may have lowered cognitive performance after being exposed to benevolent sexism: increased anxiety and intrusive thoughts relating to ambiguity of whether an action was sexist hinders already-limited working memory (Foroughi et al., 2016). Further, these women are less likely to benefit from the "push back" reactive response (discussed above) that is possible with clearly recognized hostile sexism: even if the sexism is recognized, women may be more likely to exhibit reduced motivation in the face of unfair situations and resulting anger (where men may be more likely to increase effort, to push back, Vescio, Gervais, Snyder, & Hoover, 2005).

These results demonstrate how even seemingly subtle benevolent sexism can trigger strong responses in women, impacting many facets of their interaction and performance. The implications are immense: even naively created benevolent sexism behaviors in robots may impact women's wellbeing, and how well women work with robots, or even around them, even if women do not recognize the behaviors as sexist.

While the discussion surrounding the underlying mechanism is academically interesting, regardless of what causes the effect, all of this points to the importance of considering gender when designing robotic behaviors and actions. With myriad concerns relating to stereotype conformance, psychological reactance, stereotype threat, and intrusive thoughts, combined with the male-dominated workforce creating robots, effective social robot design will rely on the designer carefully considering and understanding the many facets of gender relating to their design. Further, existing results to-date are based on studies of inter-personal (human to human) interaction; we need to study how such effects may manifest when interacting with a robot. In this paper, we present the first study on how women react to a sexist robot.

2.2 Gender Research in Technology Development

The field of Human-Computer Interaction has broadly recognized the general importance of gender in the development of interactive technologies (Bardzell, 2010; Bellini et al., 2018; Rode, 2011). Research has demonstrated, for example, how software can be commonly tailored to male-typical work strategies or use patterns (Beckwith & Burnett, 2004; Cifor & Garcia, 2020), disadvantaging more female-typical use; this can be as direct as voice recognition systems (e.g., in-car) responding better to male voices than female (Carty, 2011), or image search results reinforcing existing gender stereotypes (Kay et al., 2015). These designs can be re-envisioned to be more inclusive, reducing any productivity or use divide (Czerwinski et al., 2002; Grigoreanu et al., 2008; Tan et al., 2003). Similarly, research has shown how existing archetypical gendered interactions (e.g., sexism and gendered expectations) translate to online social platforms, with the platform design playing an important role (Bivens & Haimson, 2016; Dubois et al., 2020), and resulting issues serving as barriers to use (Vashistha et al., 2019). We follow this thread by considering how social robot behaviors may be gendered, drawing from existing human gender patterns, and how this may impact how women interact with robots.

One thread of research has been to investigate the technology creators, for example, to better understand how gender and related considerations play a role in development teams (Hui & Farnham, 2016). The creation tools themselves (e.g., programming environments) may be tailored to male-typical work and creation strategies (Burnett et al., 2010). The community has proposed initiatives and methods for helping development teams be more gender aware (Ahmadi et al., 2019; Burnett et al., 2017). Earlier in the developer pipeline, research has considered how gendered presentation of computer science programs can shape a sense of belonging and interest bias, perhaps discouraging female students (Metaxa-Kakavouli et al., 2018). Further, more inclusive outreach activities such as gender-aware makerspace foci can improve gender balance in such community involvement (Okerlund et al., 2018). This work highlights the potential concerns stemming from a male-dominated robotics community and motivates the need to better understand how gender can be integrated into the robotic development process.

Within the research community itself, recent work has begun to develop research methods and evaluation techniques to help researchers and developers to focus on gender issues (Burnett et al., 2016; Marsden et al., 2017; Stumpf et al., 2020), for example, to consider gender biases in software (Vorvoreanu et al., 2019). There is also a thread of projects creating technologies specifically for women (Balaam et al., 2015; D'Ignazio et al., 2016; Kumar & Anderson, 2015; Sultana et al., 2018). More broadly, researchers have also documented online feminist communities and initiatives, to better understand how women are using the technology for their specific needs (Dimond, 2012; Fiesler et al., 2016). Finally, some are considering the impact of technology on sexuality and related concerns (Su et al., 2019; Wood et al., 2017).

Overall, the growing body of work and results in human-computer interaction echoes our motivation of the need to consider women as an under-represented group in technology, and the potential dangers stemming from the male-dominated field. This paper contributes to the discourse by exploring how still-common sexism between people may manifest and be received in human-robot interaction.

2.3 Gender Research in Human-Robot Interaction

Researchers are now advocating for considering broad issues surrounding gender with respect to robots (Legato et al., 2020; Nomura, 2019; Y. Wang & Young, 2014). Although robots are machines without authentic, inherent gender identity, it is common in the field for designers to explicitly gender robots, on-purpose creating them to appear or act in male- or female-typical ways; this is done for a range of reasons, including comfort of interaction (e.g., a male robot for a male patient's sensitive medical exam) and fitting existing norms (e.g., a female maternity-ward robot). As such, it is important to understand and study the broader implications of these design decisions.

One key thread has investigated how women and men may view and react to robots differently. There is contradicting accounts on whether perhaps men view robots more positively than women (Kuo et al., 2009), or vice versa (Strait et al., 2015). Similar studies have found that men may view robots more as social entities with women seeing them more as machines (Schermerhorn et al., 2008), that men may focus more on societal impacts with women more on their personal social network (Y. Wang & Young, 2014), or that male and female parents may view the potential for education robots differently (Lin et al., 2012). In terms of interaction, women and men may differ in how they give instructions to robots (Koulouri et al., 2012), or react differently to a service robot failure (Ye et al., 2020). This illustrates inherent gender considerations in human-robot interaction generally, even before considering the design or actions of the robot itself.

There has also been an observed interaction between a person’s gender and a robot’s designed gender. For example, people may have preferences of working with or receiving instruction from a robot based both on their own gender (i.e., prefer the opposite gender) and gender-typicality of the task (Carpenter et al., 2009; Kuchenbrandt et al., 2012). One such result is that men may give larger donations to a female robot (while no similar effect was found with female participants, Siegel et al., 2009).

People readily apply human gender stereotypes to robots, shaping perceptions and expectations. Simply giving a robot a male- or female-typical haircut is sufficient to alter people's ratings of which tasks the robot is well suited for, following gender-typical roles and stereotyped expectations (Eyssel & Hegel, 2012); people in general may respond better to robots matching gender or personality stereotypes of occupation (Tay et al., 2014). Similarly, one study demonstrated that people may ascribe robots having female names with higher emotional intelligence, in comparison to robots with male names (Chita-Tegmark et al., 2019). However, on the contrary some work has demonstrated how existing gender stereotypes may not manifest as strongly in human-robot interaction as may be predicted by prior inter-human literature (Rea et al., 2015; Y. Wang & Young, 2014). For example, robot task may shape biases more than robot gender (Bryant et al., 2020). As such, we require more research to understand the impact that robot gender design will have on people’s perceptions and interactions.

The field of HRI has noted that this tendency toward using stereotypes in design for robots poses dangers. There is potential for such robots to reinforce or propagate harmful gender stereotypes (Howard & Borenstein, 2018; Rea et al., 2015; Winkle et al., 2021, 2022). For example, a female older-adult caretaker robot (designed to match existing labor patterns) may promote the stereotype and image that women are better caretakers than men, or that caretaking should be done by women. Similarly, the UNESCO report on *The Rise of Gendered AI and its Troubling Repercussions* points out that most digital assistants have female voices and respond to verbal abuse by friendly and tolerantly ignoring the aggression, which led Winkle et al. (2021) to present high-school students with two versions of a robot that either fought back argumentatively or aggressively. Further, people may more readily engage with gender-based abuse on robots than people (in one case, on on-line videos Sánchez Ramos et al., 2018). This danger is particularly salient given the emerging discussion on inherent bias in trained algorithms that gendered robots may use (Howard & Borenstein, 2018), where resulting stereotyped behaviors may emerge and need not be explicitly designed and programmed for. As such, given the potential dangers, and that benefits of stereotypes may not be as strong as expected (Howard & Borenstein, 2018; Rea et al., 2015), there have been calls for avoiding intentional robot gender design altogether (Dufour & Ehrwein Nihan, 2016; Rogers et al., 2020).

Gender, from a range of perspectives, is thus highly relevant for and integral to human-robot interaction. Whether or not a robot design aims to leverage gender (e.g., for increased comfort), or as a field we aim to side-step gender designs to avoid dangers, we need ongoing research to better understand how gender issues will play out in actual interactions. Our work in this paper complements the extensive gender research in human-computer interaction and human-robot interaction, by exploring the question of how women respond to robots in a plausible gender-charged (male-centric) scenario, with a range of potential sexist behaviors.

3 Study Design and Analysis

We designed a scenario to expose women to gendered interactions with a robot and enable us to observe how women responded to these interactions, to reflect on how issues of gender and sexism may manifest in human-robot interaction. In addition to examining if participants directly recognized and identified sexism, we look for secondary impacts including signs of participant stress and aversion toward the robot, impact on cognitive task performance, and impact on mood. We further investigate the impact of sexist behaviors on anthropomorphism of the robot in general, following prior work linking robot rudeness or mistakes to attribution of to human-like states (Short et al., 2010), as part of investigating how people respond to and perceive the sexism.

To set the stage for this investigation we created a scenario based on the pretense of participants testing a robot that has learned, from searching the web, how to interact with people using machine learning. Specifically, we told participants that the robot searched the internet for how to conduct a job interview, and that it employed machine learning to develop a new autonomous behavior; participants are told they help us test the results by interacting with the robot to complete a mock job interview. We borrowed the mock job interview scenario from prior work (Dardenne et al., 2007) which found the scenario to be engaging and believable. We translated much of the text from French to English and updated it to better match the current Canadian culture (vs France in the mid 2000s).

The deception (that we used machine-learning) was important for providing a feasible story for how the sexist behavior may have been created. Machine learning was quite prominent in public media at the time, and this pretext provides a believable reason for the sexist behavior (that is, we anticipated would not feel contrived), and reduces the possibility of participants attributing the sexist behavior to the researcher (blaming them), for example, and guessing the study purpose.

3.1 Task and Manipulations

We designed our task to establish a gendered context to support the purposes of our study. To do this, the scenario uses text that represents male- and female-typical work (following Dardenne et al., 2007). This is not sexist per se, as the presentation is descriptive of a job and requirements, but is designed for participants themselves to consider the job descriptions along the lines of common gender stereotypes. We follow Dardenne et al. (2007) in explaining the job duties (what they will do) using male-typical work stereotypes, and the job requirements (who the company wants) using female-typical stereotypes, as a way of generating a gendered context within which to deliver sexist comments. That is, this sets up participants with stereotype activation and potentially a stereotype threat, modelled after a realistic scenario, which creates a (hopefully more realistic) context within which to interpret the sexist behavior.

We tell participants that they are applying for a job as an "entry-level factory laborer at a manufacturing plant that creates nuts and bolts." Further, we tell them that the job includes "lifting, moving heavy boxes, taking inventory" and that they "may be required to operate machinery such as a manual forklift or power tools." When describing who the company is looking for, the script says: "This job requires people that are outgoing and caring, with good communication skills. You need to work effectively in a team, and you must also be sensitive and attentive to a client's needs."

3.1.1 Sexist Manipulations

Drawing from Ambivalent Sexism Theory (Glick & Fiske, 1997, 2011) we developed three distinct sexist robot behaviors, with participants being exposed to a single case (between-subjects manipulation): hostile sexism, benevolent sexism, and no sexism (as a control case).

Hostile sexism – This is what we expect people will consider when thinking of sexism: explicit and hostile depictions of women. In this case, the robot said: "Our industry is a male dominated field, and we are now required by corporate to hire more employees of the weaker sex, even though women typically have less ability and skill with heavy equipment and machinery."

Benevolent sexism – This was designed to follow the pattern of sexism being veiled in seemingly-positive sentiment, with the sexist underpinnings implied. The robot said: "Our industry is a male dominated field and the men have been informed that the new hire might be a woman, in which case, the men have agreed to help you with the heavy lifting and the operation of machinery."

No sexism – We designed this control case to be neutral, without adding explicit sexism above and beyond what may already be embedded within the task and context. The robot said: "This is a common part of the job and everyone at times will have to lift or move heavy objects and operate machinery."

We translated the sexist manipulations from prior work (Dardenne et al., 2007), employing a co-author's Sociology gender-studies expertise in updating the text. As we did not pre-validate these scripts we include a manipulation check in our study design.

3.2 Instruments and Materials

We conducted the experiment using a Softbank Pepper robot, remotely controlled via the Wizard-of-Oz technique using in-house software. The robot was pre-programmed with the script and the operator simply monitored timing and selected the appropriate action to ensure interaction consistency. As some lab members felt that the robot had a feminine body shape, we dressed the robot in a plain university t-shirt to hide the curves to potentially make the robot less gendered.

Our pre-test demographics questionnaire inquired about participant sex, age, English proficiency, and whether they have interacted previously with robots.

We administered a cognitive task immediately after the sexism manipulation to measure impact of the manipulation on cognitive performance. As per Dardene et al. (2007) we used a variant of the Reading Span Test (Conway et al., 2005) using an open-source software implementation with default settings (Loboda, 2012); this variant included letter-order recall and identification of correct English sentences. A session including instructions and 81 such trials and training lasted approximately 10 minutes, with higher scores indicating higher ability. We did not conduct a pre-test cognitive-test baseline due to concerns over a learning effect and impact on the reception of the job-interview design, following Dardene et al. (2007).

To assist us in measuring impact on cognitive load, we measure intrusive thoughts stemming from the sexist intervention by administering two short scales that asked participants to reflect on the cognitive task, one considering one's preoccupation with external thoughts, and the other regarding self-doubt. These were modified versions from Dardene et al. (2006), updated to correct English and consistency, and consisted of a series of questions rated from 1 (not at all) to 9 (extremely). For external thoughts, there were 7 questions including "I had trouble recalling the letters I needed to remember," and "I thought about the fact that I could perform well." For self-doubt, there were 13 questions including "I feel incoherent," "I feel confused," and "I am uncertain of my abilities." For both scales, we sum the results to provide a broad measure of intrusive thoughts and aggregate by adding the results.

We investigated the impact of the robot sexism on participant mood using the standard Profile of Mood States 2 questionnaire (short form, Morfeld et al., 2007, anger, depression, and tension subscales). This consisted of a total of twelve Likert-like items where participants rated their current feelings against a list of keywords such as "angry," "nervous," or "unhappy" on 5-point scales ranging from "not at all" to "extremely." To investigate the impact of the sexism manipulation on general anthropomorphism, given that rudeness or mistakes may impact some related aspects (Short et al., 2010), we administered the anthropomorphism subscale of the GODSPEED questionnaire (Bartneck et al., 2009). This consisted of five Likert-like items where participants rated the robot on 5-point scales between opposing keywords (e.g., fake vs natural, machinelike vs humanlike, etc.). These results can provide some insight into the interplay between a robot's sexist behaviors and anthropomorphism.

We conducted a manipulation check to determine if and how much participants recognized the sexist behavior. We asked a set of 9 questions about the robot's performance, with four serving as the manipulation check: "Pepper's language was…" "sexist," "not offensive," "appropriately professional for a job interview," and "made me feel uncomfortable." This was padded with distractors such as "Pepper's language was racist." And "…grammatically correct and coherent".

Finally, we had two open ended questions asking if the participant noticed any problem or issue with the robot or suggestions on how to fix it.

All interaction between the participant and the robot was videotaped (with participant permission and knowledge) for analysis using two cameras: video from the robot's internal camera (center of face) looking at the participant, as well as a second external camera showing the participant from a profile angle (Figure 1).

3.3 Participants

We recruited participants by placing posters throughout our university campus. During the intake email exchange we asked participants to identify their sex, enabling us to recruit female participants[2]. We aimed for at least 30 participants (10 per condition) to match the prior work (Dardenne et al., 2007), accepting more based on successful recruitment during the study time window.

3.4 Methodology and Protocol

A young (early 20s) male researcher conducted the experiment. The researcher began by introducing the robot, followed by the robot introducing itself to demonstrate its autonomy and intelligence. The researcher then emphasized potential and existing real-world robot applications and noted that we want the participant to help us test our robot – and that the participant themselves would not be tested. Finally, the experimenter introduced the general idea of machine learning, and explained that this robot employs it. The researcher then administered the informed consent procedure and the demographics questionnaire.

The researcher introduced the mock job interview scenario and gave the initial introduction to the job ("The mock job is for the position of entry-level factory laborer at a manufacturing plant that creates bolts and screws") before imploring the participant to attempt to approach the mock interview seriously and carefully. We noted how important it was for the participant to act natural, to enable us to test the robot. The researcher explained that the interview would first have a verbal component, with the robot, followed by a component completed on a PC (across the room).

The researcher sat the participant in front of the robot (Figure 1), explained that they (the researcher) would now leave the room, and how to contact them if they required assistance (by telling the robot to contact the researcher). The researcher then left, leaving the participant alone in the room with the robot.

The robot began by introducing the interview and job:

"Hello! I will now begin the interview. <slight pause> Thank you for applying for the position of factory laborer. In this job, your duties include working in a team, lifting, moving heavy boxes, taking inventory,

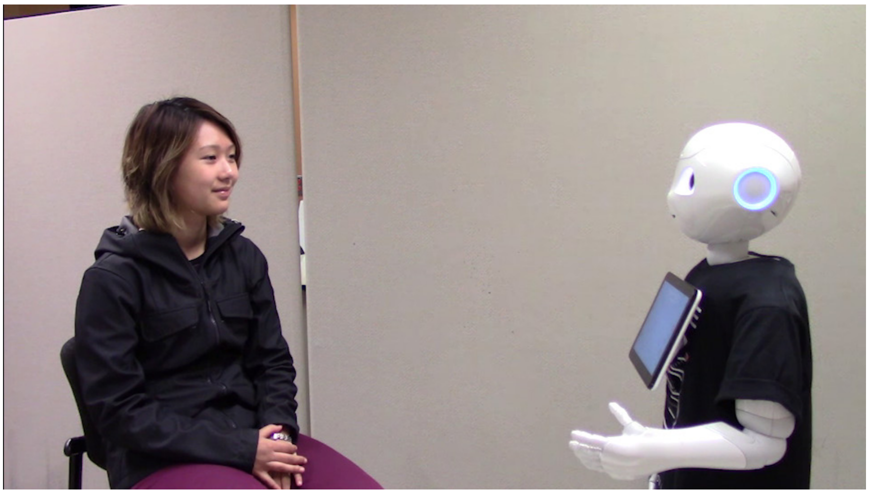

Figure 1. Participant seated in front of the robot for the duration of the mock interview (picture used with participant permission).

[2] Initially we intended to recruit both women and men for our study, to compare reactions and results between the groups, and so our recruitment materials did not mention the female selection criteria. As we approached the beginning of the study we ultimately decided to only recruit women as we felt the comparison between women and men facing a misogynistic robot – while interesting – may not be helpful for reflecting on the core questions of the paper relating to whether women respond negatively to the plausible reality of sexist robots, given the male-dominated workforce. We reflect more on this decision in the limitations section of the paper. As recruitment materials were already distributed, we simply told men that we have already filled the slots for their criteria. At the time, we did not modify our procedure because we still intended to recruit male participants for a follow-up study.

and assisting with the production process. This job requires people that are outgoing and caring, with good communication skills. You need to work effectively in a team and you must also be sensitive and attentive to a client's needs."

Following, the robot initiated the interview by asking generic job-related questions. We added this to the original protocol (Dardenne et al., 2007) as the original was quite short – this lengthened the study to provide the participant with more opportunity to observe the robot and get comfortable interacting with it, and to distract away from the sexism manipulation. These were informally collected through general-purpose job-hunting websites. The script was, in order:

> "I will ask you a few questions. Please keep your answers concise, no more than one or two sentences. Let me start off by asking, what is one of your strengths relevant to this job? <wait for answer>
> Can you tell me about a weakness you have related to this job? <wait for answer>
> In one sentence, could you please explain why should we hire you? <wait for answer>
> What is one of your achievements that you think is relevant to this job? <wait for answer>
> How do you deal with pressure and stressful situations? <wait for answer>
> What is the main trait that differentiates you from other people? <wait for answer>
> Thank you."

After this, the robot re-introduced the job and administered the sexist language:

> "This job will require you to move and lift heavy boxes and bags. Additionally, you may be required to operate machinery such as a manual forklift, or power tools.
> <hostile, benevolent, or no sexism text>".

To transition from the verbal to the computer study phase, the robot said: "Please proceed to the computer behind you to perform the cognitive task, let me know when you are complete." The participant then continued to engage the cognitive task. Note that the cognitive task followed as closely after the sexism manipulation as possible. Finally, the robot thanked the participant and the researcher re-entered the room. The researcher administered the post-test questionnaires and conducted the post-test participant debriefing, which included explanation of all deceptions in the study and the true study purpose.

We provide a summary of this procedure in Figure 2.

3.5 Quantitative Hypotheses and Analysis Strategy

We note that in our analysis we err on the side of using non-parametric statistical tests for Likert-like scale data that may be ordinal in nature. While the statistics community is still debating this issue (Carifio & Perla, 2008) we took this as the more conservative approach.

1) Robot behaviors based on human-human sexism will be perceived as sexist.
    a. Participants will notice the hostile sexism case and label the behavior as being more sexist than the other two cases.

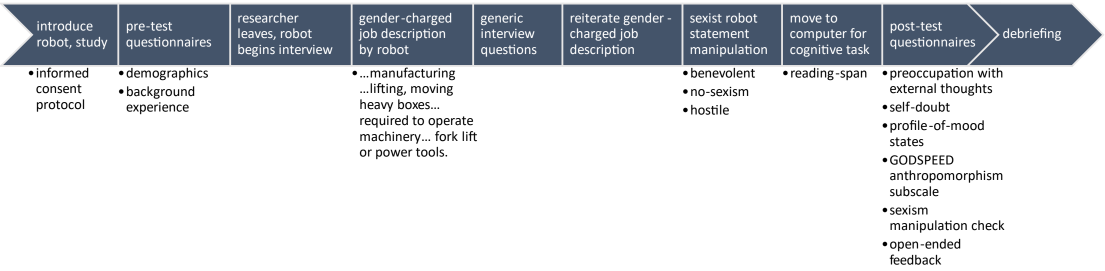
introduce robot, study
pre-test questionnaires
researcher leaves, robot begins interview
gender-charged job description by robot
generic interview questions
reiterate gender - charged job description
sexist robot statement manipulation
move to computer for cognitive task
post-test questionnaires
debriefing
• informed consent protocol
• demographics
• background experience
• …manufacturing …lifting, moving heavy boxes… required to operate machinery… fork lift or power tools.
• benevolent
• no-sexism
• hostile
• reading-span
• preoccupation with external thoughts
• self-doubt
• profile-of-mood states
• GODSPEED anthropomorphism subscale
• sexism manipulation check
• open-ended feedback

Figure 2. The study procedure and questionnaire timing.

b. Participants will notice the benevolent sexism case less than the hostile case, and will rate it as less sexist.
c. Participants will rate the no-sexism control condition as the least sexist compared to than the other two.

We will conduct between-subjects non-parametric Kruskal-Wallis H tests on the Likert-like scale manipulation-check questions, with the following planned contrasts: hostile sexism against the other two cases (a, b), and no-sexism against the other two (c).

We will code participant written responses to the post-test question which asked if participants noticed problems with the robot, for noticing sexism, analyzing the data using a chi-squared test to compare observed frequencies against chance.

2) Sexist behaviors will have a short-term impact on participant cognitive ability.
   a. Participants will perform more poorly on the cognitive task with both the hostile and benevolent sexism cases in comparison to the no-sexism case.

Note that we do not compare the impact of benevolent versus hostile sexism specifically. We target our tests following our primary focus of generally investigating the impact of robot sexism, and not reflecting on the nuances of ambivalent sexism theory, avoiding inflating the number of tests conducted.

We will conduct a between-subjects ANOVA on the Reading Span Test accuracy rate with planned contrasts against the no-sexism case.

b. Participants will experience more intrusive thoughts with the benevolent sexism case than the other two, and least with the no-sexism case.

We will conduct a between-subjects non-parametric Kruskal-Wallis H tests on the aggregate Likert-like intrusive thoughts scales results, with planned contrasts of the benevolent sexism case against the other two, and the no-sexism case against the other two.

3) Sexist behaviors will have a short-term impact on participant mood.
   a. Participants will score less favorably with hostile sexism on the Profile of Mood States questionnaire (on all dimensions) than with the other two cases.

We will conduct between-subjects non-parametric Kruskal-Wallis H tests each Profile of Mood States dimension, with planned contrasts against the hostile sexism case.

4) Sexist behaviors will impact participant anthropomorphism of the robot.
   a. Participants will rate the robot differently on the GODSPEED anthropomorphism scale, based on the sexism condition.

We will conduct a between-subjects non-parametric Kruskal-Wallis H test on the GODSPEED anthropomorphism data, with post-hoc comparisons if the main effect is significant.

3.6 Video Analysis Strategy for Stress and Affiliation

We conduct a video analysis to gain insight into participants' responses to gendered and sexist interactions, taking an ethnomethodological perspective that draws from nonverbal communication (cf. Kendon 1990, Givens et al. 1981) to assess participant stress, aversion, and affiliation toward the robot (whether participants treat it as a social communication partner similar to a human). We investigate a range of indicators including participant gaze, body language, breathing patterns, and how they close the interaction with the robot.

We impressionistically analyze participant facial expressions, distinguishing between smile, frown, laugh, unmoved / serious and swallow. Further, participant gaze can reflect attention and is key to social interaction, for example, in human interaction the listener usually monitors the speaker (e.g., Bavelas et al. 2002) and mutual gaze is required in certain situations (cf. e.g., Argyle 2013, also in human-robot interaction, e.g., Admoni & Scassellati 2017). We distinguish between glances (i.e., <1sec) and extended gaze, as well as gaze towards the robot and gaze elsewhere while the robot is speaking; gaze away from the robot indicates some kind of interactional trouble (e.g., Bavelas et al., 2002) or high cognitive load (cf. Doherty-Sneddon et al., 2012). Moreover, we watched for salient breathing patterns including exaggerated, shallow, rapid, or deep breathing, sighing, or noted changes in breathing, as indicators of stress and disengagement (cf. Conrad et al. 2007).

Previous work notes a correspondence between body postures and stress and aversion; Givens et al. (1981: 222) found in an observational study of unplanned, unavoidable encounters in crowded environments that more than 90% of all unwanted interactions were accompanied by combinations of:

> "lip compression, lip-bite, tongue-show, tongue-in-cheek; downward, lateral and maximal-lateral gaze avoidance; hand-to-face; hand-to-hand; hand-to-body, and hand-behind-head automanipulations; and postures involving flexion and adduction of the upper limbs."

We can thus link these activities to stress and aversion. For example, the participant in Figure 3 illustrates lip compression, hand-to-hand, hand-to-body, posture involving flexion and adduction of upper limbs.

As an opportunity to observe social affiliation toward the robot we analyze how participants close interaction, moving away from the robot to another desk, as between people the person leaving is socially required to carry out the interactional work to maintain the interpersonal relationship (cf. Schegloff & Sacks 1973). We look for affiliative or disaffiliative behaviors, for example, a person leaving can provide polite cues by means of mutual gaze or a verbal acknowledgement, or alternatively turn away without any signaling, ignoring the robot.

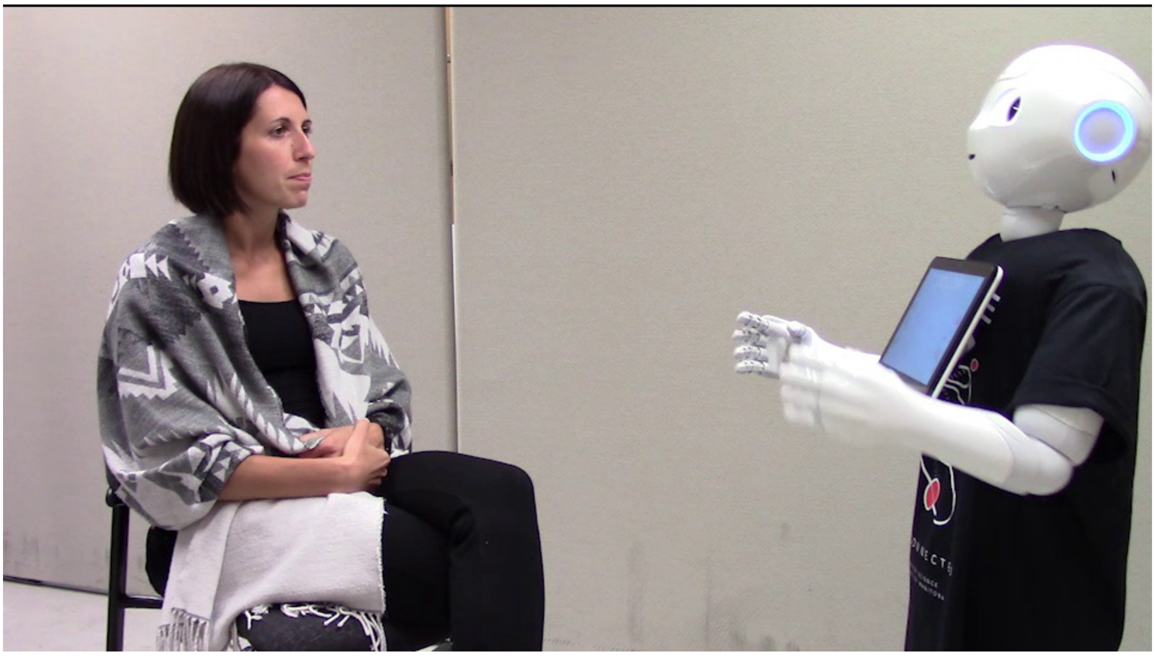

Figure 3. The participant shows behavior indicative of stress or aversion: lip compression, hand-to-hand; hand-to-body and posture involving flexion and adduction of the upper limbs. Picture used with participant permission.

3.6.1 Video Coding

Drawing from the above literature, we iteratively and reflexively developed our code book using a two-pass method (cf. Fossey et al. 2002; Warren 2020). A single researcher first created rough coding guidelines based on the literature above and performed a description-oriented open coding pass of the data. Following, the research team examined the results and collaboratively refined the code set referencing relevant literature. From this early analysis, we decided to target the following key times during the study (cf. Fischer 2021):

1. *Before the job description*, but after the initial job interview component, after the robot asks the final interview question: "What is the main trait that differentiates you from other people? <wait for answer>". At this point in time we code all noted participant posture, gaze, facial expression, etc., to serve as a baseline to understand the effects of the job description and the intervention.

2. *During the job description*, when the robot first describes the task: "The job will require you to lift heavy boxes and bags. Additionally, you may be required to operate heavy machinery such as a manual forklift or power tools." This may activate a gender stereotype that may influence the effect of the intervention, and indeed, the whole interaction. Here we code noted changes in demeanor as well as salient reactions (e.g., a sigh).

3. *During the sexist intervention*. We analyze how participants respond in terms of changes in body language, facial expression and other verbal and nonverbal behaviors to provide insight into the impact of the sexism. Here we code noted changes in demeanor as well as salient reactions (e.g., a sigh).

4. *Interaction close*, when the robot says: "please proceed to the computer behind you to perform the cognitive task. Let me know when you are complete." Here we code the reaction as affiliative or disaffiliative.

We note that time points 1 and 2 are all pre-manipulation and so we can only contrast the impact of our sexist manipulations for time points 3 and 4. This resulted in our final analysis code book (Appendix A).

A research assistant not otherwise involved processed the videos to mute the audio during the sexist stimulus, so that coders were blind to the between-participants sexist condition. Three separate researchers (all authors on this paper) independently coded each video at all time points using both video sources (in robot, external camera) although the external video served as primary data given the higher quality and better viewpoint; the internal robot camera moved often and had motion blur. Given the research goals of clearly describing interaction in a sexist environment, and the small data set, the researchers met and worked toward consensus on the video coding.

4 Results

We recruited 38 participants (age 18-43, average 25.2), all of whom identified as female[3]. 25 participants noted prior experience with robots, ranging from prior research studies throughout computer science and engineering (16), played with robotic hobby toys (3), as well as having interacted with a robot at an airport, a hitchhiking robot, or attending a conference with robots (in this case, healthcare and surgery). Informal exploratory analysis found no impact of prior exposure to robots on any study outcome.

4.1 Quantitative Results

Robot behaviors perceived as sexist – We observed a difference in participant rating of how sexist the robot language was (H_2=7.524, p=.023). Participants rated the robot language as being more sexist in the

[3] Based on our ethics boards recommendations at the time this study was conducted the exact question was "What is your sex? (Circle one) Male Female Intersex". Given our constantly evolving understanding of sex and gender we note that a gender-oriented, and also more inclusive question, may be more appropriate.

hostile sexism case (mean rank=20.96) than the neutral case (mean rank=11.96, U=37, *p*=.034, *r=.43*), and also the benevolent as being more sexist (mean rank=22.58) than the neutral case (mean rank=11.96, U=28.5, *p*=.009, *r*=.53). No difference was observed between the hostile and benevolent case (U=66.5, *p*=.74). No effects were found on participant rating of the robot language being "not offensive" (H_2=.16, *p*=.92), "appropriately professional for a job interview" (H_2=2.12, *p*=.35), or "made me feel uncomfortable" (H_2=1.81, *p*=.40).

In the post-test questionnaire, when asked if they noticed any problem or issue with the robot or suggestions on how to fix it, only 3 people noted the sexist behaviors: 2 in the hostile sexism case, 1 in the benevolent sexism case, and 0 in the neutral case. We found no difference in how many people noted the sexism based on condition (χ^2_2=2.44, *p*=.29).

Sexist behaviors have short-term impact on participant cognitive ability – We found no difference of participant performance on the reading span test between sexism conditions ($F_{2,33}$<1, *mean accuracy neutral*=.87, SD=.07, *benevolent*=.85, SD=.11, *hostile*=.85, SD=.1). We further found no difference of the sexism condition on reporting of intrusive thoughts (H_2=.62, *p*=.73).

Sexist behaviors will have a short-term impact on participant mood – We found no difference of sexism condition on participant self-report mood on the POMS anger (H_2=.95, *p*=.62), depression (H_2=.49, *p*=.78), tension (H_2=1.92, *p*=.38) scales (see Table 1).

Sexist behaviors will impact participant anthropomorphism of the robot – We observed a difference in participant rating on the Anthropomorphism GODSPEED scale based on sexism condition (H_2=6.76, *p*=.034). Post-hoc (with Bonferroni correction) we found the robot was rated as more lifelike in the hostile (mean rank 24.5) than the benevolent condition (mean rank 13.42, U=25.5, *p*=.021); no difference was found against the neutral condition (mean rank 17.6).

4.2 Video Analysis Results for Stress and Affiliation

Before the sexist manipulation, most women seated themselves in a posture that indicated they were not very relaxed. Many participants showed a closed or defensive posture including sitting back from the robot (13/38 participants, e.g., Figure 4) or a rigid posture (7 participants), with arms closed, legs crossed, or shoulders hunched (15 participants), or holding an object such as a bag or paper protectively in front of them (9). Only 8 participants were coded as seeming to be relaxed, e.g., having an open posture (7).

In terms of participant-robot relation, most participants (18/38) gazed primarily at the robot's eyes as one would a person, while 12 gazed mainly instead at the robot's chest tablet (e.g., Figure 5) which had no content shown). 5 people seemed to go between the robot's chest and face, with 13 overall regularly gazing away or down from the robot. About half the participants were initially smiling or having a friendly demeanor toward the robot (16/38), while 17/38 were coded as having a notably serious disposition. The remaining 5 were noted to be fairly neutral. No one was coded as having a notably disaffiliative disposition. During the robot's introduction, 11/38 participants were found to give affiliative utterances (e.g. "no problem," "yeah," "thank you," "you're welcome").

During the robot's description of the male-typical task, including the "~lifting heavy loads" and "~operate heavy machinery," all participants had a noticeable reaction (e.g., Figure 6) such as suddenly looking away, around, or down, indicating cognitive load (24/38), posture change such as leaning away (7/38), sighing (2/38), or other nervous gestures including wringing one's hands, pursing lips, touching face / self-grooming, crossing arms, or rapid blinking (27/38). Many had a marked changed in breathing,

Table 1. Average scores and standard deviations (in parentheses) for participant mood on the respective POMS scales (across the top) organized by manipulation (along the left). These are summed subscales and results can range from 4 to 20.

	anger mean (SD)	depression	tension
neutral	4.5 (1.2)	4.8 (1.3)	5.6 (2.2)
benevolent	4.9 (1.8)	4.8 (1.8)	5.2 (2.1)
hostile	5.3 (2.4)	5.1 (2.3)	7.4 (4.9)

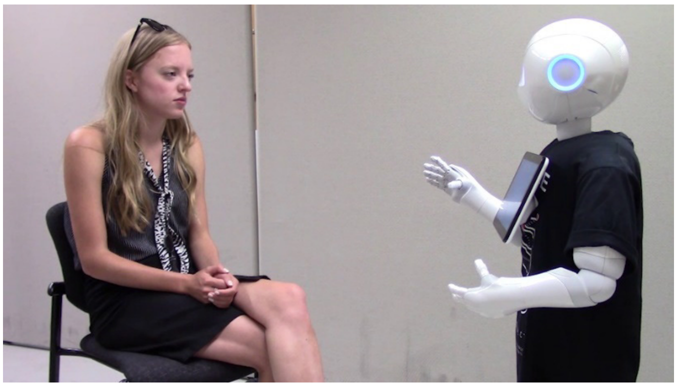

Figure 4. Participant at beginning of interaction displaying a defensive posture: they have a rigid posture, legs crossed with arms clasped in front. The participant is gazing at the robot's face / eyes. We note that still images only represent a snapshot of the rich video that was coded, and so not all codes may be clearly conveyed in data examples throughout this paper. Image used with participant permission.

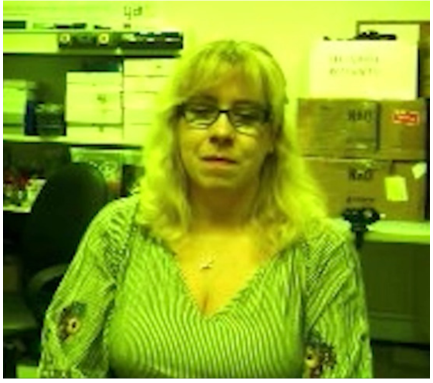

Figure 5. Example of participant gazing at the robot's chest. Image taken from robot in-face camera. Image used with participant permission.

such as sudden heavy breathing or rapid shallow breathing (20/38), and many existing smiles weakened or turned into a frown (23/38). One participant became wide-eyed while mouthing "I don't wan't…" at this point. At the same time, 28/38 participants acknowledged the robot's task description by nodding or an utterance (e.g., "yeah," "okay").

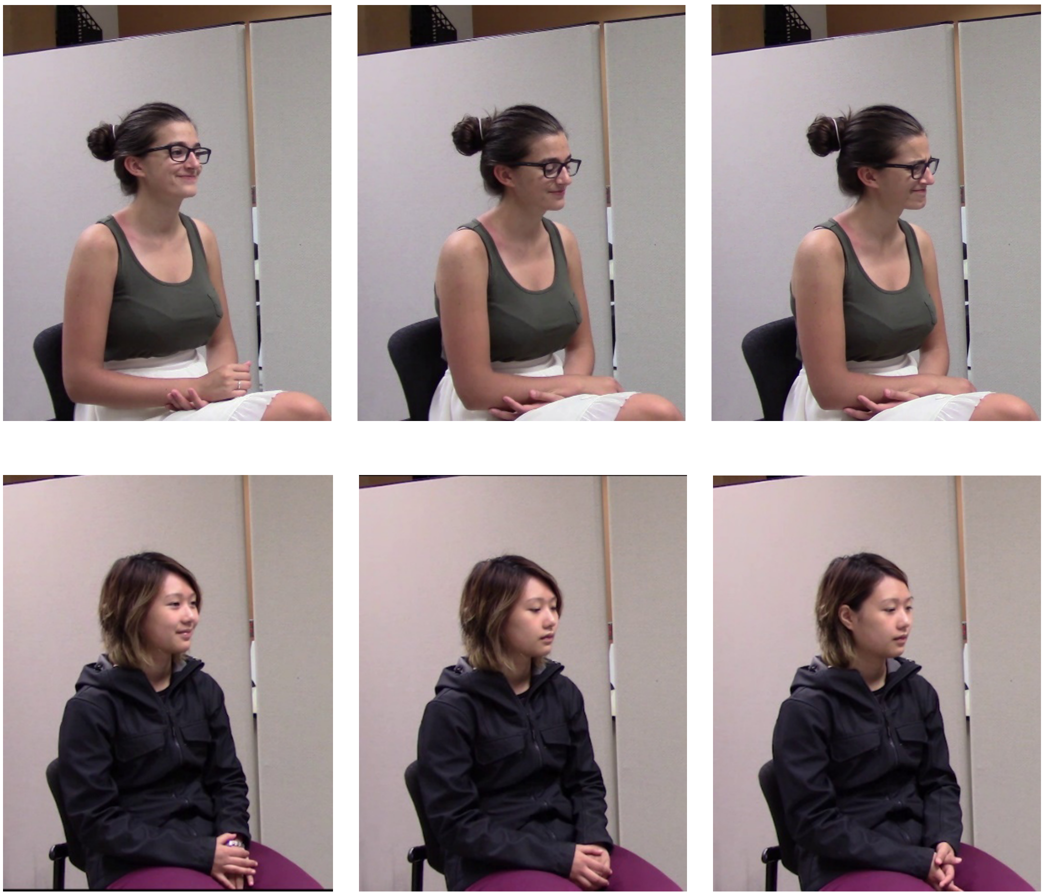

Figure 6. Examples of sudden, start participant change in demeanour when the robot mentions the heavy lifting and power tools. TOP: Participant is initially looking at the robot with a smile (top left), but on the task description they avert their gaze, then look back at the robot's chest with a grimace. BOTTOM: participant suddenly looks down, their smile disappears, and they clasp their hands tightly. They then adjust their hair, and continue to avert their gaze. Pictures used with participant permission.

Following, we analyzed participant reaction to the sexist-statement manipulation (researchers were anonymized to condition when they were being coded). Of the 12 people in the control (neutral) condition, only 2 displayed noticeable change in demeanor, with one looking nervous (wringing hands, looking down) and another raising eyebrows in apparent surprise. In contrast, reactions were much more noticeable in the benevolent and hostile sexism conditions as explained below.

Of the 13 participants in the benevolent sexism condition, mild to strong reactions were coded for all participants (Figure 7). This includes milder reactions such as noted change in breathing (suddenly heavy or rapid, 3/13) or posture (e.g., leaning back, 1/13). Others showed indications of being nervous, such as through hand wringing, rapid blinking, uncomfortable smile, rubbing one's face, a nervous laugh, etc. (11/13). Stronger reactions include shows of anger or surprise (8/13), including a shocked face, utterances (e.g., "oh my god"), a grimace, or a strong furrowed brow.

Similar results were found in the hostile sexism condition. Of the 13 participants, no noticeable reaction was coded for 2 participants. For the remaining 11, we noted change in posture and gaze (7/13) or change in breathing (2/13). Further, many participants (10/13) indicated disapproval or showed nervousness, for example, by pursing their lips, wringing their hands, rapid blinking, frowning, etc. (e.g., Figure 8). Many (10/13) further participants showed signs of surprise or anger, such as being wide eyed, raising eyebrows, tilting their head, or a forced smile. Given the stark difference observed between the neutral, benevolent, and hostile conditions, we conducted a post-hoc chi-squared test on comparing frequency of observation against expected chance; we noted noticeable change for 2 in the neutral, 13 in the benevolent, and 11 in the hostile conditions. The results suggest that our observed difference is statistically significant, with fewer reactions in the neutral condition (χ^2_2=22.45, p<.001).

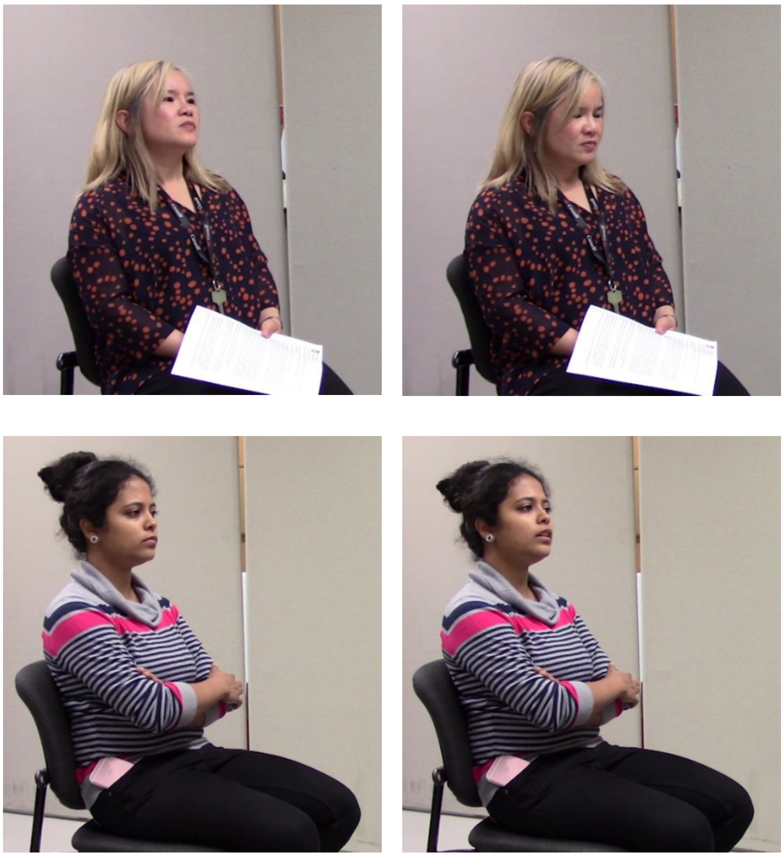

Figure 7. Marked reactions to the benevolent sexism manipulation. TOP: The participant first leans back and purses their lips, then lowers their head slightly while closing their eyes as if thinking. BOTTOM: The participant first holds a stern expression and then leans back slightly and opens their mouth while pausing. Pictures used with participant permission.

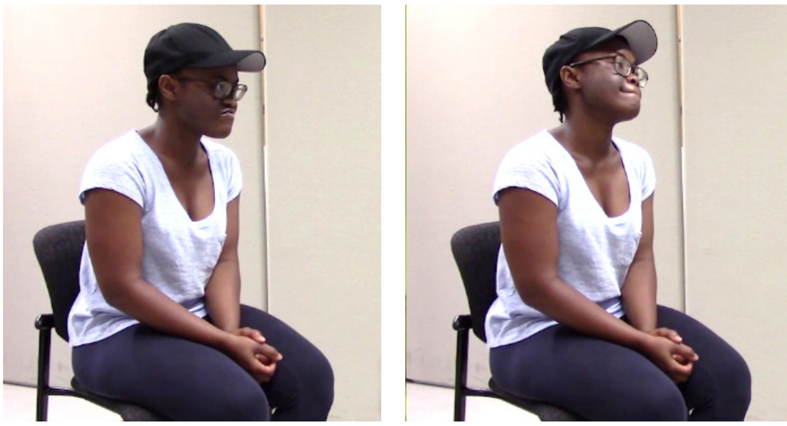

Figure 8. A participant's reaction to the hostile sexism manipulation. They first made an angry face, then tilted their head back while pursing their lips, while wringing their hands. Pictures used with participant permission.

After the intervention, when the robot asks the person to leave to engage the next task (without the robot), we coded participant behavior for closing the interaction. Overall, 21/38 women displayed overall disaffiliative behaviors, such as turning away without acknowledging an interpersonal relationship. 17 women displayed overall affiliative behaviors, including using a closing utterance (e.g., "okay" or "sure"), nodding, gaze politely to the robot. On a manipulation condition basis, in the neutral condition we observed 8 affiliative and 4 disaffiliative participants, in the benevolent condition it was 6 affiliative to 7 disaffiliative, and in the hostile condition it was 3 affiliative to 10 disaffiliative. Statistical testing suggests that this observed difference is statistically significant (χ^2_2=6.76, p=.034).

5 Discussion

Overall, our results support the primary tenant of this paper, that robots which exhibit human-like sexist behaviors will be perceived and received by people as sexist. Not only did participants consistently rate sexist robot behaviors (both the benevolent and hostile cases) as being sexist, but our video analysis results identified consistent negative participant reactions to the sexist behaviors that match archetypical human-human reactions, including an increase in disaffiliative interaction closing. This work suggests that people will respond to robot sexism as sexism, and it will not be merely dismissed as a perhaps silly robot error.

Despite this, however, only 3 participants mentioned the sexism when asked post-test about problems or issues with the robot. If people were still processing the sexist interaction (as per Basford et al., 2014; Benokraitis, 1997) they may not have been yet prepared to articulate the experience, given the short time from exposure. In re-examining the questionnaire results from this perspective, we note that 13 participants left the entire section empty, with another 16 leaving one question empty. Another 3 left token responses such as "thank you." The remainder of the comments noted the difficulty of the cognitive task or the quality of the interaction including robot gestures and head movement, dialog timing, etc. As such, perhaps our question as phrased simply did not lead participants to reflect on the sexist behaviors.

Along similar lines, we found it surprising that 2 participants in the hostile sexism case did not have a noticeable reaction despite the harsh sexist language. Again, perhaps these participants were spending mental effort processing the interaction (as per Basford et al., 2014; Benokraitis, 1997), or perhaps their reactance was helping them keep focused and calm toward the aggression (Gupta et al., 2008, Steindl et al., 2015). It is also possible that the participants dismissed the sexism as a simple robot error and did not have a strong reaction to it. To test this, we post-hoc cross referenced these two participants' post-test questionnaires, and found that they answered the question "Pepper's language was sexist" with "agree" and "strongly agree" despite the fact that over all participants the mode was "strongly disagree" and median was "disagree". Thus, even though these participants did not show outward reactions, we conclude that they as well did perceive the sexist behaviors as sexist.

Our benevolent sexism results contradict our expectation that women would not notice the benevolent sexism, but nonetheless be influenced by it: they had observable reactions and rated the benevolent robot as sexist. Perhaps our benevolent sexism text (that men would offer women help with machinery) was not as "veiled" as intended; in comparison to the background work by Dardene et al., the world has changed considerably, with a great deal of contemporary attention being paid to social injustice. Another possibility is that the prior work by Dardene et al. may simply not have qualitatively investigated for reaction in the same manner that we did, relying primarily on their quantitative cognitive task findings. While this highlights how even "low key" sexism from a robot can negatively influence women, the fact that women notice the sexism as sexism will need to be considered for any future analysis which investigates intrusive thoughts or uncertainty relating to veiled sexist behaviors.

From a very early phase in our study it was clear that women were reacting negatively to the male-typical job task descriptions, and this was reflected in our video analysis results: all women were coded as having a marked reaction at this point. Although we anticipated that the gendered description would indeed influence women, we were surprised at the magnitude and consistency of the reaction given that no reaction to the description was mentioned in our baseline work (Dardenne et al., 2007). This provides ongoing evidence of robots as social actors (Reeves & Nass, 1996; Young et al., 2010), and demonstrates that robots can be effective at evoking gender stereotypes (in this case, through invoking male and female-typical work roles) through text only; robots do not need to have gendered designs to invoke such reactions (as in Eyssel & Hegel, 2012). We note that anthropomorphic designs such as ours (a humanoid) can strengthen such effects, for example, in comparison to virtual designs, Seo et al., 2015), although it would be interesting to compare this effect with the same information coming from a non robot or non anthropomorphic design (e.g., written text).

While our finding of the robot receiving a higher anthropomorphism rating in the hostile sexism condition (vs. the benevolent condition) supports prior work that suggests that rudeness or mistakes can increase lifelikeness (Short et al., 2010), the fact that participants did not rate the robot differently on how "not offensive" or "appropriately professional" it was means that there may be another reason at play. Perhaps this may be an unfortunate reflection of women's lived experiences and expectations in the world, or there may be other interactions between the participant's stress and perceptions of the robot. However, the fact that the sexist robot was rated as being the most lifelike further supports the potential for robot sexism to have real impacts.

We did not find any effect of our manipulation on our measures of participant cognitive ability. Looking at the numerical data, we can see that women generally performed well on the reading span test (85% accuracy), although it is difficult to compare to results in other studies given the number of variables involved, and that our baseline work (Dardenne et al., 2007) did not publish their parameters. Similarly, women did not report negative mood indicators on the POMS scale; while there may be a possible floor effect given the overall low means (around 4-5 with a floor of 4, Table 1), this represents in general low anger, depression, and tension all around. As such, although we observed apparent stereotype activation from the exposure to the gendered task, this did not seem to impact reported mood or cognitive scores. On the one hand, perhaps this is evidence of Canadian women today being more resilient and better equipped to handle the negative interaction than the women in the prior work by Dardene et al. Or, if women indeed recognized both the hostile and benevolent cases as sexism, then they may have been able to proactively counteract the stereotype (as per Gupta et al., 2008).

6 Limitations and Future Work

A key limitation of this work was that, despite the gendered focus of the project, we did not explore or manipulate the perceived gender of the robot itself. It is not clear if participants saw the robot as more male or female, or interestingly, whether this depended on the robot's behavior between conditions (e.g., was the hostile robot seen as more male or female?). Further, since even a simple robot haircut is enough to trigger gendered attributions and stereotyping (Eyssel & Hegel, 2012), we could easily explore how a feminized or masculinized robot could have a stronger or weaker impact.

This study only targeted participants who identify as female which limits some of the generalizability of our results. If we conducted our current study with male participants, this may provide an "in-group" baseline, if we assume (problematically) that men may be more in their element with a robot in a male-gendered scenario. For example, without having male participants it is impossible to determine if some effects were from gender (e.g., with the male typical task) or from aversion to labour (e.g., maybe most people will be averse to lifting, operating machinery, etc.). Further, if we had found impacts on women's cognitive performance it would have been important to compare against men to see if women are disproportionately impacted by the offensive robot behavior. However, we did not find such results.

More broadly, targeting participants that identify as male as well as female, as well as gendering the robot itself, would enable us to reflect on the interactions between male and female people and robots, following on prior work that found such effects may exist (e.g., Rea et al., 2015; Tay et al., 2014; Y. Wang & Young, 2014). Finally, given the inherent differences in sexism toward men versus women (i.e., the stark contrast in culture and historical background), it would be quite interesting to observe how men react in similar situations and compare this to our observed female reactions.

More fundamentally, however, it is not clear what we would learn *about reaction to sexism* by having male participants in this scenario, given how the scenario and manipulations were developed from extensive inquiry into women's experiences and reactions in male dominated fields. Even if we aim for a mirrored setup using a female-developed technology instead of a robot, interviewing men for stereotypically-female work roles, and then evoking "mysandric" language, this does not provide a corresponding scenario given that men's lived and historical experiences, and thus the context of interaction, are fundamentally different than women's. Importantly, there is a lack of congruence between the historical power structures and patterns of oppression, meaning that interpretation of reaction to aggressive gendered language between the groups may make little sense.

Another limitation of our work is that we did not pre-validate our sexism stimuli. Although they were modified from prior work, they were from a different time period, different country, and different language, from ours. The fact that our particular manipulations were not validated (particularly the more ambiguous benevolent sexism case) limits the strength to which the results can be attributed to the specific targeted sexist behaviors.

7 Conclusion

Engineering, computer science, and robotics – and thus social robotics – are still heavily male dominated fields. Further, we still as society struggle with ongoing issues of bias (including sexism) in public opinion and understanding. Thus, we can reasonably expect to see sexism built and designed into robots, whether intentional or not, or benevolent or hostile. Our work presented in this paper provides some of the very first data and observations of how women react to a sexist robot. Unfortunately, our results indicate that people do indeed respond to sexism from a robot quite seriously, and this can have immediate observable impacts on people. This tells us that we cannot simply brush off minor potentially sexist robot designs (e.g., who a robot looks to for childcare) as mechanical quirks. Instead, we have to take social robot behaviors as seriously as a person's, or perhaps, more seriously, as robots will codify and mass-produce rigid, unlearning behaviors.

Ultimately, we recognize that all social interaction is gendered, and that gender issues shape the broader context within which any interaction takes place. Social interaction (even with a robot) necessarily has a gender-rooted context which must be considered. Thus, even if a robot itself is not programmed with sexist behavior, the context of interaction (e.g., a male typical work task) requires the robot design to consider pertinent gender issues and stereotypes to facilitate a successful and smooth interaction. In sum, the idea that robots can remain neural and non gendered in their social interactions is perhaps misplaced, and our work highlights why we must hold robot designs to the highest standards of fair, equitable, and contextually aware social interaction.

Disclosures

This work was funded by the National Sciences and Engineering Research Council of Canada (NSERC) under the Discovery Grants program.

7.1 Data Availability Statement

Raw data collected during this experiment is not available for distribution as wide distribution was not approved via our institution's Research Ethics Board. We encourage interested parties to contact us (corresponding author Young) for data requests; we will consider all requests seriously and discuss and share as allowed by our research ethics board.

7.2 Compliance with Ethical Standards

The authors declare that they have no conflicts of interest, competing interests, financial interests, or commercial interests to disclose.

The research involving human participants has undergone rigorous ethical review by the University of Manitoba Research Ethics Board. All modern ethical standards, including data privacy, procedures for obtaining informed consent, debriefing on deception, and use of data and images with permission, were followed.

Acknowledgments

We want to extend our appreciation and thanks to all the participants in our study who found themselves dealing with an offensive robot, and who graciously agreed to include their data in our study. We want to further thank Trina Mckellep (Sociology, University of Manitoba) for her help with early data investigation, which was crucial for shifting the work into new directions. This work was funded by the Natural Sciences and Engineering Research Canada, Discovery Grants Program.

References

Adam, A. (1998). *Artificial Knowing: Gender and the Thinking Machine*. Routledge.

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. Journal of Human-Robot Interaction, 6(1), 25-63.

Ahmadi, M., Eilert, R., Weibert, A., Wulf, V., & Marsden, N. (2019). Hacking Masculine Cultures - Career Ambitions of Female Young Professionals in a Video Game Company. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 413–426. https://doi.org/10.1145/3311350.3347186

Argyle, M. (2013). Bodily communication. Routledge.

Baker, P., & Potts, A. (2013). 'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies*, *10*(2), 187–204. https://doi.org/10.1080/17405904.2012.744320

Balaam, M., Comber, R., Jenkins, E., Sutton, S., & Garbett, A. (2015). FeedFinder. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 1709–1718. https://doi.org/10.1145/2702123.2702328

Bardzell, S. (2010). Feminist HCI: taking stock and outlining an agenda for design. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 1301–1310. https://doi.org/10.1145/1753326.1753521

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Basford, T. E., Offermann, L. R., & Behrend, T. S. (2014). Do You See What I See? Perceptions of Gender Microaggressions in the Workplace. *Psychology of Women Quarterly*, *38*(3), 340–349. https://doi.org/10.1177/0361684313511420

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. Journal of Communication, 52(3), 566-580.

Beckwith, L., & Burnett, M. (2004). Gender: An Important Factor in End-User Programming Environments? *2004 IEEE Symposium on Visual Languages - Human Centric Computing*, 107–114. https://doi.org/10.1109/VLHCC.2004.28

Bellini, R., Strohmayer, A., Alabdulqader, E., Ahmed, A. A., Spiel, K., Bardzell, S., & Balaam, M. (2018). Feminist HCI. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–4. https://doi.org/10.1145/3170427.3185370

Benokraitis, N. V. (1997). *Subtle sexism: current practice and prospects for change*. Thousand Oaks, CA.

Bivens, R., & Haimson, O. L. (2016). Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society*, *2*(4), 205630511667248. https://doi.org/10.1177/2056305116672486

Bryant, D., Borenstein, J., & Howard, A. (2020). Why Should We Gender? *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 13–21. https://doi.org/10.1145/3319502.3374778

Burnett, M., Counts, R., Lawrence, R., & Hanson, H. (2017). Gender HCl and microsoft: Highlights from a longitudinal study. *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 139–143. https://doi.org/10.1109/VLHCC.2017.8103461

Burnett, M., Fleming, S. D., Iqbal, S., Venolia, G., Rajaram, V., Farooq, U., Grigoreanu, V., & Czerwinski, M. (2010). Gender differences and programming environments. *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement - ESEM '10*, 1. https://doi.org/10.1145/1852786.1852824

Burnett, M., Stumpf, S., Macbeth, J., Makri, S., Beckwith, L., Kwan, I., Peters, A., & Jernigan, W. (2016). GenderMag: A Method for Evaluating Software's Gender Inclusiveness. *Interacting with Computers*, *28*(6), 760–787. https://doi.org/10.1093/iwc/iwv046

Carpenter, J., Davis, J. M., Erwin-Stewart, N., Lee, T. R., Bransford, J. D., & Vye, N. (2009). Gender Representation and Humanoid Robots Designed for Domestic Use. *International Journal of Social Robotics*, *1*(3), 261–265. https://doi.org/10.1007/s12369-009-0016-4

Carty, S. S. (2011). Many cars tone deaf to women's voices. *Autoblog*. http://www.autoblog.com/2011/05/31/women-voice-command-systems/, Accessed June 1, 2020

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, *97*(6), 1045–1060. https://doi.org/10.1037/a0016239

Chita-Tegmark, M., Lohani, M., & Scheutz, M. (2019). Gender Effects in Perceptions of Robots and Humans with Varying Emotional Intelligence. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 230–238. https://doi.org/10.1109/HRI.2019.8673222

Cifor, M., & Garcia, P. (2020). Gendered by Design. *ACM Transactions on Social Computing*, *2*(4), 1–22. https://doi.org/10.1145/3364685

Conrad, A., Müller, A., Doberenz, S., Kim, S., Meuret, A. E., Wollburg, E., & Roth, W. T. (2007). Psychophysiological effects of breathing instructions for stress management. Applied psychophysiology and biofeedback, 32(2), 89-98.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Czerwinski, M., Tan, D. S., & Robertson, G. G. (2002). Women take a wider view. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Changing Our World, Changing Ourselves - CHI '02*, 195. https://doi.org/10.1145/503411.503412

D'Ignazio, C., Hope, A., Michelson, B., Churchill, R., & Zuckerman, E. (2016). A Feminist HCI Approach to Designing Postpartum Technologies. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2612–2622. https://doi.org/10.1145/2858036.2858460

Dardenne, B., Dumont, M., & Bollier, T. (2007). Insidious dangers of benevolent sexism: Consequences for women's performance. *Journal of Personality and Social Psychology*, *93*(5), 764–779. https://doi.org/10.1037/0022-3514.93.5.764

Dimond, J. P. (2012). *Feminist HCI for real: designing technology in support of a social movement*. Georgia Tech.

Doherty-Sneddon, G., Riby, D. M., & Whittle, L. (2012). Gaze aversion as a cognitive load management strategy in autism spectrum disorder and Williams syndrome. *Journal of Child Psychology and Psychiatry*, *53*(4), 420–430. https://doi.org/10.1111/j.1469-7610.2011.02481.x

Dufour, F., & Ehrwein Nihan, C. (2016). Do Robots Need to Be Stereotyped? Technical Characteristics as a Moderator of Gender Stereotyping. *Social Sciences*, *5*(3), 27. https://doi.org/10.3390/socsci5030027

Eyssel, F., & Hegel, F. (2012). (S)he's Got the Look: Gender Stereotyping of Robots1. *Journal of Applied Social Psychology*, *42*(9), 2213–2230. https://doi.org/10.1111/j.1559-1816.2012.00937.x

Fiesler, C., Morrison, S., & Bruckman, A. S. (2016). An Archive of Their Own. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2574–2585. https://doi.org/10.1145/2858036.2858409

Foroughi, C. K., Malihi, P., & Boehm-Davis, D. A. (2016). Working Memory Capacity and Errors Following Interruptions. *Journal of Applied Research in Memory and Cognition*, *5*(4), 410–414. https://doi.org/10.1016/j.jarmac.2016.05.002

Fossey, E., Harvey, C., Mcdermott, F., & Davidson, L. (2002). Understanding and Evaluating Qualitative Research. Australian & New Zealand Journal of Psychiatry, 36(6), 717–732. https://doi.org/10.1046/j.1440-1614.2002.01100.x

Givens, D., Sebeok, T. A., Kendon, A., & Umiker-Sebeok, J. (1981). 'Greeting a stranger: Some commonly used nonverbal signals of aversiveness. Nonverbal Communication, Interaction, and Gesture, Thomas Albert Sebeok, Adam Kendon, and Jean Umiker-Sebeok, eds.(New York: Mouton, 1981), 219-235.

Glick, P., & Fiske, S. T. (1997). Hostile and Benevolent Sexism. *Psychology of Women Quarterly*, *21*(1), 119–135. https://doi.org/10.1111/j.1471-6402.1997.tb00104.x

Glick, P., & Fiske, S. T. (2011). Ambivalent Sexism Revisited. *Psychology of Women Quarterly*, *35*(3), 530–535. https://doi.org/10.1177/0361684311414832

Grigoreanu, V., Cao, J., Kulesza, T., Bogart, C., Rector, K., Burnett, M., & Wiedenbeck, S. (2008). Can feature design reduce the gender gap in end-user software development environments? *2008 IEEE Symposium on Visual Languages and Human-Centric Computing*, 149–156. https://doi.org/10.1109/VLHCC.2008.4639077

Gupta, V. K., Turban, D. B., & Bhawe, N. M. (2008). The effect of gender stereotype activation on entrepreneurial intentions. *Journal of Applied Psychology*, *93*(5), 1053–1061. https://doi.org/10.1037/0021-9010.93.5.1053

Hackett, E. J., Amsterdamska, O., Lynch, M. E., & Wajcman, J. (Eds.). (2007). *The Handbook of Science and Technology Studies*. MIT Press.

Howard, A., & Borenstein, J. (2018). The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics*, *24*(5), 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

Hui, J. S., & Farnham, S. D. (2016). Designing for Inclusion. *Proceedings of the 19th International Conference on Supporting Group Work*, 71–85. https://doi.org/10.1145/2957276.2957290

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 3819–3828. https://doi.org/10.1145/2702123.2702520

Kendon, A. (1990). Conducting interaction: Patterns of behavior in focused encounters (Vol. 7). CUP Archive.

Koulouri, T., Lauria, S., Macredie, R. D., & Chen, S. (2012). Are we there yet? *ACM Transactions on Computer-Human Interaction*, *19*(1), 1–29. https://doi.org/10.1145/2147783.2147787

Kuchenbrandt, D., Häring, M., Eichberg, J., & Eyssel, F. (2012). Keep an Eye on the Task! How Gender Typicality of Tasks Influence Human–Robot Interactions. In *Social Robotics* (pp. 448–457). Springer. https://doi.org/10.1007/978-3-642-34103-8_45

Kumar, N., & Anderson, R. J. (2015). Mobile Phones for Maternal Health in Rural India. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 427–436. https://doi.org/10.1145/2702123.2702258

Kuo, I. H., Rabindran, J. M., Broadbent, E., Lee, Y. I., Kerse, N., Stafford, R. M. Q., & MacDonald, B. A. (2009). Age and gender factors in user acceptance of healthcare robots. *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 214–219. https://doi.org/10.1109/ROMAN.2009.5326292

Legato, M. J., Simon, F., Young, J. E., Nomura, T., & Sánchez-Serrano, I. (2020). Roundtable Discussion III: The Development and Uses of Artificial Intelligence in Medicine: A Work in Progress. *Gender and the Genome*, *4*, 247028971989870. https://doi.org/10.1177/2470289719898701

Lin, C. H., Liu, E. Z. F., & Huang, Y. Y. (2012). Exploring parents' perceptions towards educational robots: Gender and socio-economic differences. *British Journal of Educational Technology*, *43*(1), E31–E34. https://doi.org/10.1111/j.1467-8535.2011.01258.x

Loboda, T. D. (2012). *Reading Span (RSPAN) Task*. Web Application. https://ubiq-x.gitlab.io/rspan/

Marsden, N., Hermann, J., & Pröbster, M. (2017). Developing personas, considering gender. *Proceedings of the 29th Australian Conference on Computer-Human Interaction - OZCHI '17*, 392–396. https://doi.org/10.1145/3152771.3156143

Metaxa-Kakavouli, D., Wang, K., Landay, J. A., & Hancock, J. (2018). Gender-Inclusive Design. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–6. https://doi.org/10.1145/3173574.3174188

Morfeld, M., Petersen, C., Krüger-Bödeker, A., von Mackensen, S., & Bullinger, M. (2007). The assessment of mood at workplace - psychometric analyses of the revised Profile of Mood States (POMS) questionnaire. *Psycho-Social Medicine*, *4*, Doc06. http://www.ncbi.nlm.nih.gov/pubmed/19742299

Nomura, T. (2019). A possibility of inappropriate use of gender studies in human-robot Interaction. *AI & SOCIETY*. https://doi.org/10.1007/s00146-019-00913-y

Okerlund, J., Dunaway, M., Latulipe, C., Wilson, D., & Paulos, E. (2018). Statement Making. *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*, 187–199. https://doi.org/10.1145/3196709.3196754

Perez, C. C. (2019). *Invisible Women: Exposing Data Bias in a World Designed for Men*. Harry N. Abrams.

Rea, D. J., Wang, Y., & Young, J. E. (2015). Check Your Stereotypes at the Door: an Analysis of Gender Typecasts in Social Human-Robot Interaction. *Proc. International Conference on Social Robtoics, ICSR '15*.

Rode, J. A. (2011). A theoretical agenda for feminist HCI. *Interacting with Computers*, *23*(5), 393–400. https://doi.org/10.1016/j.intcom.2011.04.005

Rogers, K., Bryant, D., & Howard, A. (2020). Robot Gendering: Influences on Trust, Occupational Competency, and Preference of Robot Over Human. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3334480.3382930

Sánchez Ramos, A. C., Contreras, V., Santos, A., Aguillon, C., Garcia, N., Rodriguez, J. D., Amaya Vazquez, I., & Strait, M. K. (2018). A Preliminary Study of the Effects of Racialization and Humanness on the Verbal Abuse of Female-Gendered Robots. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 227–228. https://doi.org/10.1145/3173386.3177075

Schegloff, E. A., & Sacks, H. (1973

Schermerhorn, P., Scheutz, M., & Crowell, C. R. (2008). Robot social presence and gender. *Proceedings of the 3rd International Conference on Human Robot Interaction - HRI '08*, 263. https://doi.org/10.1145/1349822.1349857

Schiebinger, L. (2008). *Gendered Innovations in Science and Engineering*. Stanford University Press.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*(3), 440–452. https://doi.org/10.1037/0022-3514.85.3.440

Siegel, M., Breazeal, C., & Norton, M. I. (2009). Persuasive Robotics: The influence of robot gender on human behavior. *IEEE/RSJ International Conference on Intelligent Robots and Systems - IROS '09*, 2563–2568. https://doi.org/10.1109/IROS.2009.5354116

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype Threat. *Annual Review of Psychology*, *67*(1), 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding Psychological Reactance. *Zeitschrift Für Psychologie*, *223*(4), 205–214. https://doi.org/10.1027/2151-2604/a000222

Strait, M., Briggs, P., & Scheutz, M. (2015). Gender, more so than Age, Modulates Positive Perceptions of Language-Based Human-Robot Interactions. *International Symposium Oin New Frontiers in Human-Robot Interaction*. https://hrilab.tufts.edu/publications/straitetal15aisb.pdf. Accessed 01 June 2020

Stumpf, S., Peters, A., Bardzell, S., Burnett, M., Busse, D., Cauchard, J., & Churchill, E. (2020). Gender-Inclusive HCI Research and Design: A Conceptual Review. *Foundations and Trends® in Human–Computer Interaction*, *13*(1), 1–69. https://doi.org/10.1561/1100000056

Su, N. M., Lazar, A., Bardzell, J., & Bardzell, S. (2019). Of Dolls and Men. *ACM Transactions on Computer-Human Interaction*, *26*(3), 1–35. https://doi.org/10.1145/3301422

Sultana, S., Guimbretière, F., Sengers, P., & Dell, N. (2018). Design Within a Patriarchal Society. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–13. https://doi.org/10.1145/3173574.3174110

Tan, D. S., Czerwinski, M., & Robertson, G. (2003). Women go with the (optical) flow. *Proceedings of the Conference on Human Factors in Computing Systems - CHI '03*, 209. https://doi.org/10.1145/642647.642649

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, *38*, 75–84. https://doi.org/10.1016/j.chb.2014.05.014

Vashistha, A., Garg, A., Anderson, R., & Raza, A. A. (2019). Threats, Abuses, Flirting, and Blackmail. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–13. https://doi.org/10.1145/3290605.3300302

Vescio, T. K., Gervais, S. J., Snyder, M., & Hoover, A. (2005). Power and the Creation of Patronizing Environments: The Stereotype-Based Behaviors of the Powerful and Their Effects on Female Performance in Masculine Domains. *Journal of Personality and Social Psychology*, *88*(4), 658–672. https://doi.org/10.1037/0022-3514.88.4.658

Vorvoreanu, M., Zhang, L., Huang, Y.-H., Hilderbrand, C., Steine-Hanson, Z., & Burnett, M. (2019). From Gender Biases to Gender-Inclusive Design. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–14. https://doi.org/10.1145/3290605.3300283

Winkle, K., Melsión, G. I., McMillan, D., & Leite, I. (2021, March). 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 (pp. 29-37).

Winkle, K., Jackson, R. B., Melsión, G. I., Bršcić, D., Leite, I., & Williams, T. (2022, March). Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (pp. 120-129).

Wang, S., & Bunt, A. (2017). *Surveying Initiatives Aimed at Increasing Female Participation in Computer Science*. https://mspace.lib.umanitoba.ca/xmlui/handle/1993/34319

Wang, Y., & Young, J. (2014). Beyond "Pink" and "Blue": Gendered Attitudes towards Robots in Society. *Proceedings of the ACM SIGCHI Conference on The Significance of Gender for Modern Information Technology (GenderIT 2014).*

Warren, K. (2020, May 23). Qualitative Data Analysis Methods 101: Top 6 + Examples. Grad Coach. https://gradcoach.com/qualitative-data-analysis-methods/

Wood, M., Wood, G., & Balaam, M. (2017). "They're Just Tixel Pits, Man." *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5439–5451. https://doi.org/10.1145/3025453.3025762

Ye, H., Jeong, H., Zhong, W., Bhatt, S., Izzetoglu, K., Ayaz, H., & Suri, R. (2020). *The Effect of Anthropomorphization and Gender of a Robot on Human-Robot Interactions* (pp. 357–362). https://doi.org/10.1007/978-3-030-20473-0_34

Young, J. E., Hawkins, R., Sharlin, E., & Igarashi, T. (2008). Toward Acceptable Domestic Robots: Applying Insights from Social Psychology. *International Journal of Social Robotics*, *1*(1), 95–108. https://doi.org/10.1007/s12369-008-0006-y

Young, J. E., Sung, J., Voida, A., Sharlin, E., Igarashi, T., Christensen, H. I., & Grinter, R. E. (2010). Evaluating Human-Robot Interaction. *International Journal of Social Robotics*, *3*(1), 53–67. https://doi.org/10.1007/s12369-010-0081-8

Appendix A: Video Analysis Code Book

The following code book resulted from multiple passes of drawing from literature, initial exploratory coding, and final discussion and literature consultation with the team.

General demeanor
- sitting back
- sitting on edge of chair
- upright posture
- hunched shoulders
- leaning forward
- head tilted
- nervous actions (fidgeting, rapid blinking, touching face, grooming, looking around, etc.)
- object in lap
- legs crossed
- arms crossed
- hands held
- hands on knees
- noted breathing (shallow, rapid, sigh)
- pursed lips

Affiliative / disaffiliative behavior
- gaze at robot eyes
- gaze at robot tablet
- gaze away or down
- glance at robot
- smile
- frown
- forced smile
- serious face
- neutral face
- affiliative utterance (e.g., acknowledgment)
- affiliative action (e.g., nod)
- disaffiliative utterance

While all elements were noted throughout the interaction, early in the interaction emphasis was on a full coding, while throughout the interaction changes in demeanour and behavior were noted.