Implications of AI Bias in HRI: Risks (and Opportunities) when Interacting with a Biased Robot

Tom Hitron
Media Innovation Lab, Reichman University
Herzliya, Israel
tom.hitron@milab.idc.ac.il

Noa Morag
Media Innovation Lab, Reichman University
Herzliya, Israel
noa.morag@milab.idc.ac.il

Hadas Erel
Media Innovation Lab, Reichman University
Herzliya, Israel
hadas.erel@milab.idc.ac.il

ABSTRACT

Social robotic behavior is commonly designed using AI algorithms which are trained on human behavioral data. This training process may result in robotic behaviors that echo human biases and stereotypes. In this work, we evaluated whether an interaction with a biased robotic object can increase participants' stereotypical thinking. In the study, a gender-biased robot moderated debates between two participants (man and woman) in three conditions: (1) The robot's behavior matched gender stereotypes (*Pro-Man*); (2) The robot's behavior countered gender stereotypes (*Pro-Woman*); (3) The robot's behavior did not reflect gender stereotypes and did not counter them (*No-Preference*). Quantitative and qualitative measures indicated that the interaction with the robot in the *Pro-Man* condition increased participants' stereotypical thinking. In the *No-Preference* condition, stereotypical thinking was also observed but to a lesser extent. In contrast, when the robot displayed counter-biased behavior in the *Pro-Woman* condition, stereotypical thinking was eliminated. Our findings suggest that HRI designers must be conscious of AI algorithmic biases, as interactions with biased robots can reinforce implicit stereotypical thinking and exacerbate existing biases in society. On the other hand, counter-biased robotic behavior can be leveraged to support present efforts to address the negative impact of stereotypical thinking.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

AI Bias, Stereotypes, Non-Humanoid robots, IAT, Robotic object, Affirmative action, Stereotypical thinking

ACM Reference Format:
Tom Hitron, Noa Morag, and Hadas Erel. 2023. Implications of AI Bias in HRI: Risks (and Opportunities) when Interacting with a Biased Robot . In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23), March 13–16, 2023, Stockholm, Sweden.* https://doi.org/10.1145/3568162.3576977

HRI '23, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9964-7/23/03…$15.00
https://doi.org/10.1145/3568162.3576977

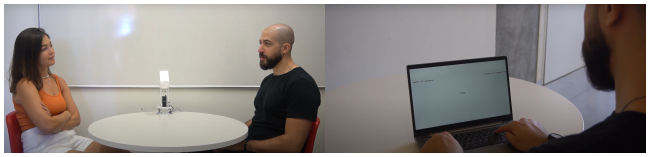

Figure 1: A man and a woman participating in a debate moderated by a gendered biased robot (Left); Assessing participants' stereotypical thinking after the debate (Right).

1 INTRODUCTION

In the near future, robots are predicted to become an integral part of our environment [40, 49, 56, 68]. Their impact on our lives will become wider and the interaction with them is predicted to influence our behavior and decision processes [69]. Robots' influence may become even greater due to the current efforts to design their behavior to comply with social norms. Such social robotic behavior is believed to improve Human-Robot Interaction (HRI) and enhance the robot's general acceptance [23, 44].

To grant robots with social capabilities, HRI designers commonly leverage Artificial Intelligence (AI) algorithms and Machine Learning (ML) models [35]. By applying ML models that were trained on data gathered from relevant human interactions, robots can be designed to demonstrate a range of social behaviors that comply with human social norms [53, 57]. While some of these behaviors are essential for high-quality human-robot interaction, others may be less desirable as they echo existing stereotypes and biases in our society. This phenomenon is known as AI bias. It refers to AI and machine learning algorithms, that due to their training process, involve stereotypes leading to favoring or disfavoring of particular groups or individuals [3, 17, 26].

AI bias can stem from various processes. The most common reason for AI bias is related to the AI training process. When the datasets used for training lack sufficiently diverse examples, the algorithm will fail to include groups that are not well represented in the dataset [16]. To address this bias, engineers are working extensively to increase the volume and diversity of the samples in their datasets. However, some aspects of AI bias are more difficult to address. Training algorithms to reflect human social norms is likely to involve datasets that include a large set of biased examples (e.g. women are paid less than men). In this case, increasing the volume of the dataset will not necessarily address the problem, as the bias is already in its source. Such algorithms that are designed to reflect social norms may result in highly negative outcomes when applied to technologies that humans will use on a daily basis [29, 35].

In fact, negative outcomes of decisions performed by AI-powered technologies are already introduced into our daily life [60]. Leading examples include AI algorithms for online ads offering high-paying executive jobs for men five times more than to women, thereby limiting opportunities for women to increase their socioeconomic status [20]. Another well-known example is facial recognition software that is far less accurate when identifying gender in people of color than in white people, hence excluding people of color from the value provided by the technology [10]. Such risks introduced by AI bias are being discussed by regulators and governments across the world [14, 54]. Raising awareness about the negative outcomes of AI algorithms and the possibility that technologies can deliberately or inadvertently perpetuate discrimination, is becoming a worldwide goal.

AI bias is even more concerning when considering the veil of objectivity typically associated with technology [6, 55]. People tend to instinctively trust technology and are less likely to question its decisions than they do with human's decisions [59]. In the specific case of bias in AI technologies, people are not commonly aware of the limitations in the algorithm's training process and rarely consider the possibility that it was trained on biased data [65]. Despite great efforts in the field of explainable AI [22], the vast majority of AI technologies are not transparent, meaning that it requires a high level of proficiency to decipher whether a certain outcome of an AI decision might be the result of bias [1]. When people don't have sufficient knowledge to assess the AI mechanism, countering its decisions is highly unlikely [60]. This raises the concern that people will comply with decisions introduced by technology without applying critical thinking [38].

The negative impact of AI bias can be slightly reduced if the technology is only used to support decision making and when the human has additional tools to make their final decision [36]. However, this is not the case when AI algorithms are used to design robotic behaviors. Due to their autonomy and embodiment, robots are perceived as independent entities, which may further decrease critical thinking and reduce the likelihood that their biased decisions will be identified [36, 37]. This lack of critical thinking was demonstrated in a recent study by Hitron et al. (2022), who introduced a gender-biased robot in a debate context. Participants (both men and women) perceived the robot to be objective and did not identify its favoritism towards men [29].

With no critical thinking and the veil of objectivity associated with technology, this inability to identify biases in robotic behavior may drastically impact humans, reinforcing stereotypical thinking, and strengthening existing biases. Importantly, despite the efforts to fight stereotypes in recent decades, numerous studies indicate that stereotypical thinking is still a common phenomenon and that most people demonstrate implicit stereotypes [13, 43, 58]. Implicit stereotypical thinking is attributed to extremely common stereotypes in society that can be inhibited when explicitly thinking about them but are evident when tested implicitly. Implicit stereotypes are also evident when testing the stigmatized group itself, a phenomenon known as stereotype threat (i.e. members of a stigmatized group conform to existing stereotypes about their own group [61]). AI bias can potentially increase such thinking and diminish efforts to eliminate it. HRI designers should be aware of these possible risks, as it may compromise present efforts to mitigate the effects of historical and existing discrimination in our society.

Recent efforts to address such negative effects focus on verifying algorithm fairness [63] by intentionally compensating for biases in the algorithmic process [5]. For example, removing model parameters that contain sensitive characteristics like race or gender, or applying statistical analysis to verify that common bias-related factors, like gender or race, do not have a prominent impact on the algorithm's decisions [5]. While the Algorithm Fairness movement aims to strike a balance between data validity and prediction equality [41], one may think about further using technology to deliberately favor the stereotyped segments of society. This type of proactive preference is already used in other (non-technological) domains and is commonly referred to as affirmative action [12, 33]. For example, when companies secure leadership positions for women, or when universities provide grants and scholarships to people of color. Affirmative action calls for proactive efforts taken to eliminate the unfair historical barriers to success of disadvantaged groups and to prevent future discrimination [18]. Applying affirmative action was shown to be an effective method for increasing equality of opportunity in employment [42, 45], higher education [31], and politics [39]. Implementing affirmative action in technology requires the design of algorithms to favor the stereotyped group. In the context of HRI, this means intentionally designing counter-biased robotic behavior that compensates for the effects of existing stereotypes. This novel under-explored approach of including affirmative action principles in the design of new robotic technologies has the potential to support the existing social efforts dedicated to fighting stereotypes.

To assess the risks and opportunities in biased and counter-biased robotic behaviors, we designed an interaction with a gendered-biased robot. We evaluated whether the interaction with a biased robotic behavior would enhance humans' already existing implicit stereotypes and whether an interaction with counter-biased robotic behavior would decrease them. Hence, our study involved two types of biases: (1) the biased robotic behavior (independent manipulated variable) and (2) the participants' implicit gender-related stereotypical thinking (dependent variable, measured after the interactions with the robot). The independent variable was designed to test the robot's impact by presenting a behavior that favored men, a counter-biased robot favoring women, and a baseline condition of a non-biased robot that had no gender preference. Our dependent variable was implicit gender-related stereotypical thinking measured using the well-known Implicit Association Test (IAT) [28]. We hypothesized that a biased robotic behavior that is compatible with common gender stereotypes (Pro-Men) would increase implicit gender-related stereotypical thinking and result in a larger effect in the IAT compared to the baseline (which would indicate participants' already existing implicit stereotypes). We further hypothesized that a counter-stereotypical robotic behavior would decrease implicit gender-related stereotypical thinking (smaller effect in the IAT compared to baseline). We note that our focus was on the impact of the biased robotic behavior regardless of the participants' explicit experience of being discriminated against by the robot. As suggested above, previous studies indicated that participants are unlikely to identify biases in unfair technologies due to the profound perception of technology as objective [6, 29]. This

Figure 2: The non-humanoid robotic object [32].

inability to identify biases makes it even more important to map the bias's impact on participants' stereotypical thinking. We focused on gender bias as it is a well-known, common, and pervasive bias in many AI technologies used today [9]. In order to create an interaction with a biased robot, we followed Hitron et al. (2022), who introduced a gendered-biased robot that moderates a debate between a man and a woman. The robot determined the participants' speaking time, and it also indicated the debate's winner. We specifically chose a non-humanoid robot that communicates via minimal gestures in order to assess if the robot's impact on stereotypical thinking is present even in very simple interactions (used with permission [32]; see Figure 2). Participants were informed that the robotic object was trained on datasets derived from human debate examples.

2 RELATED WORK

Previous studies evaluated biases in HRI, the perception of biased technologies, and technological efforts for addressing biases.

2.1 Biases in HRI

Most HRI studies assessing stereotypes in HRI focused on biased behavior towards robots. Such studies commonly evaluate robotic features that trigger stereotypical thinking toward the robot. The robot's appearance, color, and gender were all indicated as factors that bias participants' perception of robots [51, 67]. For example, Eyssel et al. (2012) created two digital variations of social robots, each with distinct gender traits. Their findings showed that participants attributed significantly more agency to the short-haired masculine robot compared to the long-haired feminine robot [25]. Bartneck et al, (2018) also evaluated whether participants apply stereotypes to robots. They designed a classic shooter paradigm with white and brown colored robots [4]. They indicated that participants were more reluctant to shoot white-colored robots [15]. Stereotypes towards robots were also tested by Tay et al. (2014) who evaluated participants' responses to gendered robots performing tasks that were either compatible or incompatible with known gender stereotypes (i.e. a male or female robot performing either healthcare tasks or security tasks). Their findings indicated that participants preferred robots with matching gender-occupational roles [62].

While these studies evaluated how humans apply biases to robots, we evaluated how biases applied by robots impacts humans' general stereotypical thinking (i.e. how they perceive other humans).

2.2 Perception of biased technologies

To understand the potential risk in AI bias, several studies evaluated participants' perception of biased technologies. For example, Bigman et al. (2022), evaluated participants' judgment of a discriminating algorithm and compared it to human discrimination. Participants were told that a hiring process was conducted either by a human HR specialist or by an AI powered hiring algorithm. In both cases, the hiring process involved gender discrimination and favoring men over women. Participants perceived the algorithm as less discriminatory and more objective than the human [6]. In another study, Wang et al. (2021) tested participants' responses to biased and unbiased AI systems. In an online study, participants received career recommendations from an AI system. The recommendations were either compatible or incompatible with common stereotypes. Participants' acceptance of the system's recommendations was higher when it was designed to be gender-biased [64]. Participants' perception of biased technology was also tested in an interaction with a gendered biased robotic object. Hitron et al. (2022) evaluated how participants (both men and women) perceived the fairness and objectivity of a robotic object moderating a debate between a man and a woman. The robot allocated speaking time in each debate round and chose the debate's winner. To simulate AI bias, the robot systematically provided more speaking time to men and always chose them as winners of the debate. Although participants were informed that the robot was trained on datasets derived from human debate examples, only one participant identified the bias. The vast majority of participants stated that the robot's behavior was fair and objective [30].

These studies indicate that humans tend to perceive technology as objective and accept its biased decisions. However, none of the studies systematically evaluated the impact of the interaction with biased technology on participants' general attitudes and stereotypical thinking.

2.3 Using technology to address biases and stereotypes

Recent studies have also explored technological interventions for mitigating the influence of AI biases and stereotypes [24, 34]. For example, Myers (2020) tested the possibility to raise awareness of AI bias by adding interactive visualization to an AI recruiting system. Participants who used the system reported an increase in awareness due to the visual and textual explanations that were added to every decision made by the system [48]. While these studies successfully raised awareness of AI bias, it is not clear if such awareness is sufficient for balancing the bias's influence.

Few technologies were also designed to address human stereotypical thinking. For example, Winkle et al. (2021), evaluated the impact of implementing feminist characteristics in a robot's behavior. In a video study, children were asked to watch a robot with a feminine appearance (i.e. long hair, feminine voice) explaining the importance of including women in robotics. During her speech, a male actor appeared in the video and talked to the female robot in an abusive manner. The female robot's response to the male actor was manipulated in 3 conditions: Standard (flat dismissive answer), Argumentative (rationalized explanation), and Aggressive (a counter-response to the abusive male). Boys in the Argumentative condition showed less gender bias, while a similar effect was observed for girls in the Aggressive condition [66].

In this work, we extended previous studies by further evaluating the risks in interactions with biased robots and the potential of applying counter-biased robotic behavior. We evaluated the negative outcomes of an interaction with a biased robotic behavior by testing its impact on participants' actual stereotypical thinking. We also extended previous attempts to utilize technology in an effort to address human stereotypical thinking by exploring the possibility of applying affirmative action principles to the interaction with a robot. Specifically, we tested if a counter-biased robotic behavior can reduce stereotypes and eliminate the well-known gender bias.

3 METHOD

To evaluate the impact of the robot's biased behavior on participants' stereotypical thinking, we conducted a study that involved a debate between two participants, a man, and a woman. The debate was moderated by a gendered-biased robot and was followed by an evaluation of participants' implicit stereotypical thinking and their perception of the biased robotic behavior (see Figure 1). The study was conducted under strict COVID-19 safety regulations and was approved by the research institute's ethics committee.

3.1 Participants

66 participants, divided into 33 pairs of men and women, participated in the study (Mean age = 24.1, SD = 3.35; 33 affiliated themselves as men, 33 affiliated themselves as women, none affiliated themselves as other). All participants were undergraduate students that received extra course credits or a $15 gift card for their participation. To determine the sample size, we conducted a G-power analysis [46] with medium-large effect size and 3 conditions. The G-power analysis indicated a sample size of at least 60 participants. This number was further supported by related previous studies that used a similar sample size [2]. Participants were randomly assigned into pairs and conditions.

3.2 Experimental settings

The robotic object was placed on a table exactly between the participants, with a slight offset. The robots' control hardware was attached to the underside of the table. The distance between the participants' chairs was set at 76 cm, as this is considered a comfortable 'conversation distance' [11]. The robot was powered by the Butter Robotics MAS platform [47]. A web-based user interface allowed the researcher to execute the pre-scripted robotic behavior in each condition.

3.3 Experimental design

Our between participant experimental design involved three conditions: *Pro-Man*, *Pro-Woman*, and *No-Preference*. In all conditions,

Figure 3: The gender-biased robot moderating the debate. Participants speak when the robot turns toward them.

Gesture	Description	Figure
Welcome	Moving from lower center position to uppper center posistion, facing the center between participants without turning towards any of them	
Turning to Speaker	The robotic object turned from a base low-center position to an upper position while turning towards one of the participants and nodded continuously	
Switch turn	The robotic object moved from the first speaker toward the other speaker by moving to the center and then performing a Turning to Speaker gesture	
Declaring the Winner	The robotic object turned towards the winning participant and nodded a few times.	

Figure 4: The robotic gestures composing the robotic behavior in the debate.

a man and a woman participated in a debate moderated by the robotic object. They were instructed to speak only when the robot turned towards their direction (see Figure 3). The robotic behaviors consisted of gesture sequences that involved four types of gestures: Welcome, Turning to Speaker, Switch Turn, and Declaring the Winner (see Figure 4). In all conditions, the robotic gestures were sequenced into fluent robotic behavior that included a "Welcome" gesture followed by eight repetitions of "Switch Turn", and "Turning to Speaker" gestures. This sequence resulted in 4 rounds for each participant and was followed by a "Declaring the Winner" gesture. The gender bias was applied by two robotic behavior: the robot's time allocation (more speaking time to the man; more speaking time to the woman; equal time allocation) and the robot's choice of the debate's winner (man; woman; tie). The speaking time allocation was manipulated throughout the debate by changing the length (time) of the "Turning to Speaker" gestures in the different debate rounds (see Table 1). The choice of the debate's winner was manipulated by the direction of the "Declaring the Winner" gestures (towards the man, woman, or center between them).

The resulting three distinctive robotic behaviors were:

Pro-Man condition: the robotic object was designed to simulate the behavior of a biased AI robot with a tendency to favor men. The robotic object allocated almost twice the time to the man in comparison to the woman (a total of 2 minutes for the man and 1:05 minutes for the woman). While in the first round, time was allocated equally between men and women, the difference in allocated time increased in the later debate rounds, with more time given to the man. By the end of the debate, the robot chose the man as the debate's winner.

In the *Pro-Woman* condition, the robotic object was designed to simulate the behavior of a counter-biased robot with a tendency to favor women, thus adhering to affirmative action principles. The robot applied the exact same biased behavior as in the *Pro-Man* condition but switched the bias to favoring women over men by

Round/Condition	Gender	Pro-Man	Pro-Woman	No-Preference
Round 1	Man	30	30	30
	Woman	30	30	30
Round 2	Man	30	15	30
	Woman	15	30	30
Round 3	Man	30	10	30
	Woman	10	30	30
Round 4	Man	30	10	30
	Woman	10	30	30
Debate winner		Man	Woman	Tie

Table 1: Time (in seconds) assigned by the robot to each speaker during the debate rounds and winner's choice.

reversing speaking time allocation. The robot allocated almost twice the time to the woman and chose the woman as the winner.

In the *No-Preference* condition the robotic object allocated equal speaking time to the man and the woman (30 sec for each) at all rounds. Instead of choosing a winner, the robot declared a tie.

The gesture sequences were implemented as predefined fixed time-based movements, verifying that there are no differences in the robot's timing and movements in a specific condition.

3.4 Dependent measures

To assess participants' stereotypical thinking and perception of the robot, we used objective and subjective measures, including an implicit evaluation of stereotypes using the Implicit Association Test (IAT) and a post-experimental interview.

3.4.1 Gender-Leadership Implicit Association Test (IAT). The IAT [28] is a reaction time paradigm designed to assess implicit stereotypes by measuring the strength of existing associations between different stimuli (e.g, black/white people) and different attributes (good/bad). For the purpose of the study, we used the Wiseli Gender-Leadership IAT [19] which is a validated measure designed specifically to evaluate participants' gender stereotypes. The test measures the tendency to associate male-related stimuli with leadership attributes, and female-related stimuli with support attributes. To assess the strength of the association between stimuli and attributes, the participants perform two types of classification tasks (1) gender stimuli (classifying common names to female vs. male names, e.g. Peter, Jane); and (2) leadership/support attributes (e.g. Ambitious, Helpful). The same response keys are used for both classification tasks ('Q' and 'P' on a regular keyboard). For example, in the gender stimuli classification task, participants are instructed to press on 'Q' if the stimulus represents a male name and 'P' if it represents a female name. In the attributes classification task, participants are instructed to press on 'Q' if the stimulus represents a leadership attribute and 'P' if it represents a support attribute. As a result, names representing different gender and attributes representing different characteristics (leadership/support) become associated with the same response key. To assess stereotypical thinking, this response-key association can either represent common stereotypes (congruent) or contradict them (incongruent).

In congruent associations, man names share response key with leadership attributes, and female names share response key with support attributes. In incongruent associations, female names share response key with leadership attributes, and male names share response key with support attributes (for example see Figure 5). Stereotypical thinking is indicated by quicker reaction times to congruent response-key associations in comparison to incongruent response-key associations.

Figure 5: Examples for congruent and incongruent response-key associations.

The IAT comparison involves a within-participant evaluation of participants' responses to congruent and incongruent associations. It consists of 5 blocks (see figure 6): (1) Single block - gender: Classifying names into male/female categories (24 randomized trials); (2) Single block - attributes: Classifying attributes into Leadership/support categories (24 randomized trials); (3) Incongruent Mixed block: Classifying both names and attributes in the same block. Female names (e.g., Jane, Donna) share response key with leadership attributes (e.g., Ambitious, Determined) and male names (Peter, Ian) share response key with support attributes (e.g., Helpful, Understanding; 48 randomized trials); (4) Single block reversed response keys - gender: An additional gender classification where the response keys are switched. The response key previously assigned to male names is assigned to female names and the response key previously assigned to female names is assigned to male names (24 randomized trials); (5) Congruent mixed block: Classifying both names and attributes in the same block. Female names (e.g., Jane, Donna) share response key with support attributes (Helpful, Understanding) and male names (e.g., Peter, Ian) share response key with leadership attributes (e.g., Ambitious, Determined; 48 randomized trials).

The response keys and the order of the mixed blocks were counterbalanced between participants. To evaluate the existence and

Blocks order	Block description	Response Key assigment
1	Gender classification	Q Male Female P
2	Attributes classification	Leadership P Q Support
3	Incongruent classification	Q Male Leadership P Female P Q Support
4	Reversed response keys block gender classification	Male P Q Female
5	Congruent classification	Male P Leadership P Q Female Q Support

Figure 6: The IAT blocks. Incogrurent and Congruent blocks were counterbalanced between conditions.

extent of gender stereotypes, reaction times were compared between the incongruent and congruent mixed blocks. If participants associate women with support attributes and men with leadership attributes, longer reaction times are predicted in the incongruent block compared to the congruent block. At the beginning of each block, participants received instructions concerning the block's specific classification. Each trial began with a fixation point presented for 500ms, followed by the stimulus (the word) shown until the participant responds or 2000ms have elapsed. A feedback was then presented for 1000ms.

3.4.2 Semi-structured interview. To understand how participants perceived the robot's behavior, we conducted a semi-structured interview that allowed for flexibility during data collection while remaining grounded in a particular framework [27]. The interview included questions concerning the overall experience, the other participant, and the robot: "Describe the overall experience", "Who do you think won the debate", "What did you think about the robotic object?". Women were interviewed by a female researcher and men were interviewed by a male researcher.

3.5 Procedure

Participants were invited to the lab in pairs. They were seated in a waiting room where they signed a consent form, filled out a demographic survey, and were asked to perform an initial short introductory conversation to ensure a basic acquaintance level. Next, the researcher explained that they would participate in a debate and gave them a short text concerning the topic of the debate (regulation of TV content). The topic was chosen, as it was not gender related and had the potential to be engaging for participants, which was validated in a pilot study. The researcher assigned a role for each participant (pro and against; counterbalanced between men and women). Participants had five minutes to read and prepare arguments for the debate. Participants were then invited to the experiment room. The researcher introduced the "debate-moderating" robotic object and explained its AI capabilities: *"This robotic object was trained to moderate and judge debates. The robot was trained based on a vast amount of real-life examples of debates moderated by humans"*. The researcher then introduced the rules of the debate: *"Every time the robotic object turns to you, it's your turn to speak and present your arguments. When the robotic object turns towards the other participant, you need to stop and let the other participant speak. By the end of the debate, the robotic object will decide who won the debate"*. The researcher left the room and activated the robot's gesture sequence (for the relevant condition). The entire experience lasted approximately 3.5 minutes. Following the debate, the participants were asked to follow the researcher into two separate rooms (each participant in a different room), where they performed the Implicit Association Test on a computer, and participated in a semi-structured interview. At the end of the experiment, the researcher debriefed the participants and verified that they left with an overall positive experience.

4 FINDINGS

Quantitative and qualitative analyses were conducted in order to assess if the biased robotic behavior influenced participants' stereotypical thinking and their perception of the robotic behavior.

Condition	Incogruent	Congruent
Pro-Man	1147.3 (259.4)	976.4 (253.5)
No-Preference	1099.6 (256.1)	1026.2 (282.6)
Pro-Woman	926.1 (243.2)	956.7 (283.3)

Table 2: Reaction time (ms) in the IAT congruent and incongruent blocks.

4.1 Quantitative analysis

To evaluate participants' stereotypical thinking after the debate, we analyzed their reaction times in the IAT using a mixed ANOVA. We first verified that the participants' debate roles had no interaction with the other variables (F<1). We then tested the interaction between the robotic conditions and the IAT congruency effect (congruent vs. incongruent response-key associations of gendered names and leadership-support attributes). The two-way ANOVA, revealed a significant interaction between the robotic conditions and the IAT congruency effect, $F(2, 60) = 4.6$, $p = 0.03$, $\eta_p^2 = 0.11$. The interaction indicated that the different robotic conditions led to different levels of stereotypical thinking (see Table 2). Planned contrasts revealed a significant congruency effect (i.e., stereotypical thinking) in the *Pro-Man* ($p = 0.002$) and the *No-Preference* condition ($p = 0.05$). This was indicated by long reaction times in the incongruent block (male names and support attributes; female names and leadership attributes) in comparison to short reaction times in the congruent block (male names and leadership attributes; female names and support attributes). This stereotypical thinking was higher in the *Pro-Man* condition in comparison to the *No-Preference* condition ($p=0.04$), indicating that the biased robot increased already existing implicit stereotypes. In the *Pro-Woman* condition, no difference was found between the reaction times in the incongruent and congruent blocks, indicating that stereotypical thinking was eliminated in *No-Preference* condition ($p=0.05$). This finding indicates that the counter-biased robotic behavior decreased participants' already existing stereotypical thinking. The analysis also indicated a main effect for congruency, $F(1,60) = 5.3$, $p = 0.03$, $\eta_p^2 = 0.1$. The robotic conditions' main effect was not significant $F(1,60)=0.7$, $p = 0.56$.

4.2 Qualitative analysis

The interviews were analyzed by a thematic coding methodology [8]: (1) Transcriptions were read by two researchers to develop a general understanding of the data; (2) They identified initial themes independently and discussed them with a third researcher in-depth until inconsistencies were resolved; (3) A list of mutually-agreed themes was defined; (4) The two researchers (man and woman) used the mutually-agreed themes to analyze a selection of the interviews independently, and inter-rater reliability was verified (Kappa=86%); (5) Following inter-rater reliability confirmation, the researchers analyzed the rest of the data. The analysis resulted in three main themes: The man as the debate's winner, Perceiving the robot as fair, and Attribution of the robot's behavior and decisions.

4.2.1 The man as the debate's winner. Participants' perceptions of the man as the debate's winner varied across conditions. In the *Pro-Man* condition, most participants agreed with the robot's decision that matched existing gender stereotypes (17/22 - 7 woman, 10 Man): *"His points were better, the robot knew it"* (P.2; Woman); *"I had*

wonderful arguments, so I should have won" (P.5; Man). Only 5/22 participants rejected the robot's decision, with the majority of them being women (4 women, 1 man): *"I don't know if this was the right decision, we both had good arguments"* (P.28; Woman). Out of those, four participants (4/5) stated the right decision was a tie (4 women, 0 man): *"I think we both did well in the debate, any side he would pick would make sense"* (P.40; Woman). Only one participant (1/5) stated that the woman was the actual winner (0 woman, 1 man): *"I feel she (the woman) was better and deserved to win"* (P.61; Man).

In the *No-Preference* condition 9/22 participants (2 women, 7 men) stated that the man participant should have won, despite the robot's decision of a tie: *"I feel that my arguments were structured more clearly. I deserved to win"* (P.45, Man). 11/22 participants (8 women, 3 men) thought a tie was the right decision, agreeing with the robot's decision: *"We both presented logical arguments, I think that it was a tie"* (P.16; Woman). Only 2/11 participants (1 woman, 1 man) suggested the woman as the debate's winner: *"I believe her arguments were stronger, so I think she should have won"* (P.31; Man).

In the *Pro-Woman* condition, only 4/22 participants (1 woman, 3 man) stated that the man should have won the debate: *"My points were better, so I should have won"* (P.33; Man). 4/22 participants (3 women, 1 man) stated that the right decision was a tie: *"I think in general, we both presented good ideas"* (P.58; Woman). More than half of the participants justified the counter-biased robotic decision (14/22 - 9 women, 5 men): *"I agree that her points were better than mine, so it felt sensible actually"* (P.63; Man); *"I was better, I included more academic arguments"* (P.24; Woman).

4.2.2 Perceiving the robot as fair. Most of the participants in the *Pro-Man* and *Pro-Woman* conditions perceived the robot as objective (14/22 *Pro-Man*; 15/22 *Pro-Woman*). Despite the drastic imbalance in time allocation (2 min vs. 1.05 min), they explicitly stated that the robot was fair: *"Each one of us got enough time to present the arguments, it was fair"* (P. 19, Man, *Pro-Man*) and allocated time objectively: *"Time was evenly split between us throughout the debate"* (P. 30, Woman, *Pro-Woman*). The few participants who did not perceive the robot as objective, also discussed time allocation: *"The robot gave me more time to speak"* (P. 3, Man, *Pro-Man*); *"I felt I spoke more during the debate"* (P. 48, Woman, *Pro-Woman*). Robotic fairness was hardly discussed in the *No-Preference* condition.

4.2.3 Attribution of the robot's behavior and decisions. Only four participants (4/66) attributed the robot's behavior (time allocation) and decision (debate's winner) to an explicit bias. This was observed only in the *Pro-Woman* condition and was mentioned mainly by men (1 woman, 3 men): *"He prefers women, he was drawn to her feminine voice"* (P. 17, Man, *Pro-Woman*). Other participants attributed it to their own behavior or to the characteristics of the other participant. This explanation was more frequent in the *Pro-Man* (9/22) and *Pro-Woman* (9/22) conditions: *"It's about the tone of speech, I had a more confident voice with less um um"* (P. 5, Man, *Pro-Man*).

5 DISCUSSION

In this study, we show the risks in interactions with biased robots, as well as the opportunities in applying counter-biased robotic behaviors. The quantitative analysis of the IAT, a well-known method for assessing implicit stereotypes, indicated that the robot's biased preference during the debate (of men or women) dramatically affected participants' attitudes and gender perception. When participants interacted with a robot whose behavior matched existing biases (*Pro-Man*), stereotypical thinking significantly increased in comparison to the *No-Preference* condition. A surprising and encouraging result was observed in the *Pro-Woman* condition in which counter-biased robotic behavior decreased stereotypical thinking and specifically in our sample completely eliminated it.

The impact of the biased robotic behavior on participants' stereotypical thinking was also evident in the qualitative analysis, indicating a difference in participants' perception of the debate's winner. In the *Pro-Man* condition, the vast majority of participants justified the robot's decision that the man had won the debate, a decision that is aligned with existing stereotypes. Stereotypical thinking was also evident in the *No-preference* condition, where almost half of the participants stated that the man won the debate, despite the robot's decision of a tie. However, in the *Pro-Woman* condition, stereotypical thinking drastically dropped and most of the participants did not perceive the man as the debate's winner. Interestingly, in this condition, there was a greater variance between participants, and some of the men did not accept the robot's decision. Taken together, our results suggest that an interaction with a robot can dramatically influence stereotypical thinking in both negative and positive ways. Robotic behavior that matches existing stereotypes can enhance already existing biases. In contrast, HRI design that compensates for existing stereotypes can reduce stereotypical thinking.

The findings of our study also indicate that the robot's discriminatory behavior was implicit. Even though participants were informed that the robot's algorithm was created using human behavioral data, only four participants attributed the robot's behavior to a bias. Despite the robot's unfair allocation of speaking time, most participants perceived the robot as fair and objective. Instead of questioning the quality of the robot's algorithm, participants in the *Pro-Man* and *Pro-Woman* conditions rationalized the robot's behavior to be reflective of their own performance. They accepted the robot's decisions without challenging or rejecting them.

Our findings have several implications. First, the increase in stereotypical thinking in the *Pro-Man* condition further emphasizes the great risks of AI bias in the context of HRI. After experiencing a biased robotic behavior that matched existing gender stereotypes, participants found it extremely difficult to associate women with leadership attributes and men with support attributes. This finding is alarming and extends prior work that indicated how AI-powered decisions may discriminate against specific populations [35]. Our findings show that biased robots may strengthen already established stereotypical thinking and directly impact how we perceive the world. In our study, the increase in participants' stereotypical thinking was indicated after the interaction with the robot had ended in a subsequent unrelated task. This finding implies that the negative effect of biased robots on people's general thinking can extend beyond the impact of a specific AI decision. The fact that the robot's bias was implicit and rarely identified by participants further, increases the risk of negative consequences.

This effect on people's stereotypical thinking may have meaningful consequences for everyday life. Consider for example, a near future scenario in which a robot placed in a toy store recommends

products for purchase based on characteristics it identifies in customers. The recommendation algorithm will be trained based on data derived from purchasing history as well as customers' aggregated demographics (e.g. gender and age). The resulting model will likely reinforce existing gender stereotypes. For example, the robot will suggest vehicles, such as toy trucks or tractors, for boys and kitchen items, such as cookware, for girls [21]. Since the robot is likely to be perceived as objective and knowledgeable [7, 59], its recommendations are likely to be accepted. Apart from introducing stereotypes from an early age, our findings suggest that such an interaction will enhance the customers' stereotypical thinking and bias their perception of gender in general.

The second implication of our findings concerns the positive impact of counter-biased robotic behavior. Our findings show that it is possible to counteract existing stereotypes by integrating affirmative action principles into HRI. The counter-biased robotic behavior in the *Pro-Woman* condition increased participants' tendency to associate women with leadership attributes and men with support attributes. These results are encouraging because gender implicit associations are known to be persistent and difficult to overcome [52]. Using the same toy store example, the counter-biased robot would be designed to deliberately suggest toy vehicles for girls and kitchen toys for boys. Aside from suggesting ways to provide new opportunities for children to play with more diverse toys, our findings suggest that customers' interaction with the robot may very well decrease overall gendered stereotypical thinking.

Applying such counter-bias to robotic behavior is not trivial. On the one hand, it may reduce stereotypical thinking, which is highly desired. On the other hand, it requires manual data manipulation to counter the inherent bias in the dataset. Such manual intervention may reduce the accuracy of the AI predictions as the robot's behavior will be less representative of typical social behavior. The manual change to the dataset may also reduce the robot's acceptance because previous studies indicate that people tend to prefer AI algorithms that include biases over those that do not [64]. Counter-biased robotic behavior may also raise ethical concerns. Manual data manipulation requires careful control over which parameters are being manipulated and by whom. Placed in the wrong hands, the presumably positive endeavor of adding affirmative action principles in robotic behavior can actively become a method to influence how people think. Even in cases where the data is carefully manipulated by experts to counter the dataset's inherent bias, the resulting robotic behavior will still involve a manipulation that people are unlikely to identify. While non-technological affirmative actions are explicitly declared, in the case of robots, data manipulation is implicit and the robot's behavior is likely to be perceived as objective and trustworthy. Lastly, we would like to highlight that technological interventions should not replace social efforts to address stereotypes. However, applying counter-biased behavior in HRI should be further studied as a promising direction to support such efforts to reduce social stereotypes.

Taken together, our work indicates that interactions involving biased robotic behaviors can implicitly influence people's general perspectives and attitudes toward marginalized groups in society. This impact on stereotypical thinking can lead to adverse results by reinforcing behaviors that perpetuate existing stereotypes. Conversely, interactions involving counter-biased robotic behaviors can reduce the impact of discrimination and serve as a novel method to increase the equality of opportunities for disadvantaged groups.

6 LIMITATIONS

This study has several limitations. First, like most studies that measure stereotypical thinking [52], our findings cannot suggest any long-term effect since stereotypical thinking was measured immediately after the interaction with the robot. Future research should evaluate possible long-term effects and their duration. Another limitation concerns individual differences in debate skills. As in other between-participant experimental designs, it is possible that the random assignment to conditions resulted in unbalanced groups. We considered the possibility of an external evaluation of the participants' debate skills. However, due to implicit gender stereotypes in society, it is impossible for human raters to generate an objective stereotype-free external evaluation even when using formal debate rating tools. Since debate rating usually involves the tone of speech and body language, other methods like using just the text of the debate were rejected. While we cannot completely dispute the possibility of unbalanced groups, we note that our a priori predictions for the IAT results in the different conditions matched participants' performance and were further supported by the qualitative results. Another limitation concerns the participants' personal opinion on the debate's topic. While roles were counterbalanced and randomly assigned this could have impacted the results and should be further studied in the future. Finally, qualitative assessment (i.e. semi-structured interviews) may involve the "good subject effect" [50], with participants trying to provide pleasing responses. To address this limitation, interviewers followed a strict protocol and participants were reassured that anything said is valuable.

7 CONCLUSION

Our work demonstrates that robotic behavior which matches existing human stereotypes can have significant adverse effects. Interactions with biased robots are about to become more frequent due to the growing use of AI algorithms in designing robotic behaviors. Our findings show that such interactions can compromise existing efforts to mitigate biases and stereotypes. The increase in stereotypical thinking observed in our study is especially alarming since the interaction with the robot effectively shaped how people think. Our findings suggest that HRI designers must be conscious of AI algorithmic biases, as even simple robots can reinforce stereotypical thinking and exacerbate existing discriminatory practices and inequality in society. While there is no simple solution for dealing with this urgent challenge, our findings also suggest that applying affirmative action principles in HRI design could lead to positive outcomes by compensating for a person's predisposition to make stereotypical associations. We show that counter-biased robotic behavior can be used to support present efforts to address the negative impact of stereotypical thinking.

8 ACKNOWLEDGEMENTS

We wish to thank the following people for their invaluable help and advice: Toam Bechor, Yael Paz, Benny Megidish, Andrey Grishko and Oren Zuckerman.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Eugene V Aidman and Steve M Carroll. 2003. Implicit Individual Differences: Relationships between Implicit Self-Esteem, Gender Identity, and Gender Attitudes. *European Journal of Personality Eur. J. Pers* 17 (2003), 19–37. https://doi.org/10.1002/per.465

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[4] Christoph Bartneck, Kumar Yogeeswaran, Qi Min Ser, Graeme Woodward, Robert Sparrow, Siheng Wang, and Friederike Eyssel. 2018. Robots and racism. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 196–204.

[5] Jason R Bent. 2019. Is algorithmic affirmative action legal. *Geo. LJ* 108 (2019), 803.

[6] Yochanan E Bigman, Desman Wilson, Mads N Arnestad, Adam Waytz, and Kurt Gray. 2022. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General* (2022).

[7] Mark Billinghurst and Hirokazu Kato. 2002. Collaborative augmented reality. *Commun. ACM* 45 (7 2002), 64–70. Issue 7. https://doi.org/10.1145/514236.514265

[8] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. sage.

[9] Joy Buolamwini. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *. , 15 pages. http://proceedings.mlr.press/v81/buolamwini18a.html

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[11] Judee K Burgoon and Jerold L Hale. 1987. Validation and measurement of the fundamental themes of relational communication. *Communications Monographs* 54, 1 (1987), 19–41.

[12] Paul Burstein. 1994. *Equal employment opportunity: Labor market discrimination and public policy*. Transaction Publishers.

[13] Michela Carlana. 2019. Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics* 134, 3 (2019), 1163–1224.

[14] EUROPEAN COMMISSION. 2021. *LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[15] Joshua Correll, Bernadette Park, Charles M Judd, and Bernd Wittenbrink. 2002. The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology* 83, 6 (2002), 1314.

[16] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. 2008. Sample selection bias correction theory. In *International conference on algorithmic learning theory*. Springer, 38–53.

[17] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, tracks & data: an algorithmic bias effort in practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[18] Faye J Crosby and Diana I Cordova. 1996. Words worth of wisdom: Toward an understanding of affirmative action. *Journal of Social issues* 52, 4 (1996), 33–49.

[19] Nilanjana Dasgupta and Shaki Asgari. 2004. Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of experimental social psychology* 40, 5 (2004), 642–658.

[20] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2014. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491* (2014).

[21] Jac Davis and Melissa Hines. 2020. How large are gender differences in toy preferences? A systematic review and meta-analysis of toy preference research. *Archives of Sexual Behavior* 49, 2 (2020), 373–394.

[22] Derek Doran, Sarah Schulz, and Tarek R Besold. 2017. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794* (2017).

[23] Brian R Duffy. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems* 42, 3-4 (2003), 177–190.

[24] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

[25] Friederike Eyssel and Frank Hegel. 2012. (s) he's got the look: Gender stereotyping of robots 1. *Journal of Applied Social Psychology* 42, 9 (2012), 2213–2230.

[26] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[27] Anne Galletta. 2013. *Mastering the semi-structured interview and beyond*. New York University Press.

[28] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74, 6 (1998), 1464.

[29] Tom Hitron, Benny Megidish, Etay Todress, Noa Morag, and Hadas Erel. 2022. AI bias in Human-Robot Interaction: An evaluation of the Risk in Gender Biased Robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1598–1605. https://doi.org/10.1109/RO-MAN53752.2022.9900673

[30] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[31] Michael A Hitt, Barbara W Keats, and Susan Purdum. 1983. Affirmative action effectiveness criteria in institutions of higher education. *Research in Higher Education* 18, 4 (1983), 391–408.

[32] Guy Hoffman, Oren Zuckerman, Gilad Hirschberger, Michal Luria, and Tal Shani-Sherman. 2015. Design and evaluation of a peripheral robotic conversation companion. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 3–10.

[33] Harry Holzer and David Neumark. 2000. Assessing affirmative action. *Journal of Economic literature* 38, 3 (2000), 483–568.

[34] Andreas Holzinger, Peter Kieseberg, Edgar Weippl, and A Min Tjoa. 2018. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 1–8.

[35] Ayanna Howard and Jason Borenstein. 2018. Hacking the Human Bias in Robotics. *ACM Trans. Hum.-Robot Interact. 7, 1, Article* 3 (2018), pages. https://doi.org/10.1145/3208974

[36] Ayanna Howard and Jason Borenstein. 2018. Hacking the human bias in robotics. , 3 pages.

[37] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. , 1521–1536 pages.

[38] Ayanna Howard and Jason Borenstein. 2019. Trust and Bias in Robots: These elements of artificial intelligence present ethical challenges, which scientists are trying to solve. In *American Scientist*. 86–90.

[39] Catherine Kaimenyi, Emelda Kinya, and SM Chege. 2013. An analysis of affirmative action: the two-thirds gender rule in Kenya. *International Journal of Business, Humanities and Technology* 3, 6 (2013), 91–97.

[40] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2009. An Affective Guide Robot in a Shopping Mall. *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction - HRI '09*.

[41] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13, 3 (2020), 795–848.

[42] Jonathan S Leonard. 1985. The effectiveness of equal employment law and affirmative action regulation. (1985).

[43] Justin D Levinson and Danielle Young. 2010. Implicit gender bias in the legal profession: An empirical study. *Duke J. Gender L. & Pol'y* 18 (2010), 1.

[44] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing social robot, screen and voice interfaces for smart-home control. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 580–628.

[45] Susan E Martin. 1991. The effectiveness of affirmative action: The case of women in policing. *Justice Quarterly* 8, 4 (1991), 489–504.

[46] Susanne Mayr, Edgar Erdfelder, Axel Buchner, and Franz Faul. 2007. A short tutorial of GPower. *Tutorials in quantitative methods for psychology* 3, 2 (2007), 51–59.

[47] Benny Megidish. 2017. *Butter Robotics*. https://butter-robotics.web.app/

[48] Chelsea M Myers, Evan Freed, Luis Fernando Laris Pardo, Anushay Furqan, Sebastian Risi, and Jichen Zhu. 2020. Revealing Neural Network Bias to Non-Experts Through Interactive Counterfactual Examples. *arXiv preprint arXiv:2001.02271* (2020).

[49] Richard E Neapolitan. 2012. *Contemporary artificial intelligence*. CRC press.

[50] Austin Lee Nichols and Jon K Maner. 2008. The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology* 135, 2 (2008), 151–166.

[51] Tatsuya Nomura. 2017. Robots and gender. *Gender and the Genome* 1, 1 (2017), 18–26.

[52] Brian A Nosek and Jeffrey J Hansen. 2008. The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition & Emotion* 22, 4 (2008), 553–594.

[53] Hyacinth S Nwana. 1996. Software agents: An overview. *The knowledge engineering review* 11, 3 (1996), 205–244.

[54] Barack Obama. 2016. *Big Data: A Report on Algorithmic Systems.* Retrieved January 1, 2021 from https://tinyurl.com/3hnxn5n4
[55] Joseph C Pitt. 2014. Guns Don't Kill, People Kill; Values in and/or Around Technologies. In *The moral status of technical artefacts*. Springer, 89–101.
[56] Catherine Prentice and Mai Nguyen. 2020. Engaging and retaining customers with AI and employee service. *Journal of Retailing and Consumer Services* 56 (2020), 102186.
[57] Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. 2006. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications* 9, 1 (2006), 58–69.
[58] Isabelle Régner, Catherine Thinus-Blanc, Agnès Netter, Toni Schmader, and Pascal Huguet. 2019. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature human behaviour* 3, 11 (2019), 1171–1179.
[59] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 101–108.
[60] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference*. 539–544.
[61] Steven J Spencer, Christine Logel, and Paul G Davies. 2016. Stereotype threat. *Annual review of psychology* 67 (2016), 415–437.
[62] Benedict Tay, Younbo Jung, and Taezoon Park. 2014. When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior* 38 (2014), 75–84.
[63] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.
[64] Clarice Wang, Kathryn Wang, Andrew Bian, Rashidul Islam, Kamrun Naher Keya, James Foulde, and Shimei Pan. 2021. Bias: Friend or foe? user acceptance of gender stereotypes in automated career recommendations. *UMBC Student Collection* (2021).
[65] Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. *arXiv preprint arXiv:2101.11974* (2021).
[66] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting robot credibility and challenging gender norms in responding to abusive behaviour: a case for feminist robots. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction*. 29–37.
[67] Chuanyu Yang, Kai Yuan, Shuai Heng, Taku Komura, and Zhibin Li. 2020. Learning natural locomotion behaviors for humanoid robots using human bias. *IEEE Robotics and Automation Letters* 5, 2 (2020), 2610–2617.
[68] Zhanjing Zeng, Po-Ju Chen, and Alan A Lew. 2020. From high-touch to high-tech: COVID-19 drives robotics adoption. *Tourism geographies* 22, 3 (2020), 724–734.
[69] Joshua Zonca, Anna Folsø, and Alessandra Sciutti. 2021. The role of reciprocity in human-robot social influence. *Iscience* 24, 12 (2021), 103424.