5-2021

# Effects of Victim Gendering and Humanness on People's Responses to the Physical Abuse of Humanlike Agents

Hideki Garcia Goo
*The University of Texas Rio Grande Valley*

# EFFECTS OF VICTIM GENDERING AND HUMANNESS ON PEOPLE'S RESPONSES TO THE PHYSICAL ABUSE OF HUMANLIKE AGENTS

A Thesis

by

Hideki Garcia Goo

Submitted to the Graduate College of
The University of Texas Rio Grande Valley
In partial fulfillment of the requirements for the degree of

MASTER OF ARTS

May 2021

Major Subject: Experimental Psychology

# EFFECTS OF VICTIM GENDERING AND HUMANNESS ON PEOPLE'S RESPONSES TO THE PHYSICAL ABUSE OF HUMANLIKE AGENTS

A Thesis
by
HIDEKI GARCIA GOO

COMMITTEE MEMBERS

Dr. Jason Popan
Chair of Committee

Dr. Megan K. Strait
Committee Member

Dr. Po-Yi Chen
Committee Member

Dr. Tom Williams
Committee Member

May 2021

ABSTRACT

Garcia Goo, Hideki, Effects of Victim Gendering and Humaneness on People's Responses to the Physical Abuse of Humanlike Agents. Master of Arts (MA), May, 2021, 34 pp., 3 tables, 1 figures, references, 51 titles.

With the deployment of robots at public realms, researchers are seeing more cases of abusive disinhibition towards robots. Because robots embody gendered identities, poor navigation of antisocial dynamics may reinforce or exacerbate gender-based marginalization. Consequently, it is essential for robots to recognize and effectively head off abuse.

Given extensions of gendered biases to robotic agents, as well as associations between an agent's human likeness and the experiential capacity attributed to it, we quasi-manipulated the victim's humanness (human vs. robot) and gendering (via the inclusion of stereotypically masculine vs. feminine cues in their presentation) across four video-recorded reproductions of the interaction.

Analysis from 422 participants, each of whom watched one of the four videos, indicates that intensity of emotional distress felt by an observer is associated with their gender identification and support for social stratification, along with the victim's gendering -- further underscoring the criticality of robots' social intelligence.

# DEDICATION

The completion of my master studies could not have been completed without the love and support from my family. My father, Cuco, my mother, Narda, my sister, Nozomi, and my grandparents who have forever inspired me to not give up in my educational journey. Likewise, I am very grateful for the support of my partner, Luis, who has always encouraged me in my academic pursuits and made me smile when I felt overwhelmed.

Thank you for your love and patience.

*La finalización de mis estudios de maestría no podría haberse completado sin el amor y el apoyo de mi familia. Mi padre, Cuco, mi madre, Narda, mi hermana, Nozomi y mis abuelitos siempre me han inspirado a no rendirme en mi viaje educativo. Asimismo, estoy muy agradecido por el apoyo de mi pareja, Luis, quien siempre me ha acompañado en mis búsquedas académicas y me ha hecho sonreír cuando me he sentido abrumada.*

*Gracias por su amor y paciencia*

ACKNOWLEDGEMENTS

I will always be grateful to Dr. Megan K. Strait, member of my committee, for all her mentoring and advice. Joining her lab opened before me career paths that I was not aware of before and because of that I was able to discover how much I enjoy research, she encouraged me to pursue higher education though her patience and guidance. My thanks go to my dissertation committee members and chair: Dr. Jason Popan, Dr. Po-Yi Chen, and Dr. Tom Williams. Their input and comments on my dissertation were of great help throughout the whole process.

TABLE OF CONTENTS

Page

ABSTRACT ..... iii

DEDICATION ..... iv

ACKNOWLEDGEMENTS ..... v

TABLE OF CONTENTS ..... vi

LIST OF TABLES ..... viii

LIST OF FIGURES ..... ix

CHAPTER I. INTRODUCTION ..... 1

Related Work ..... 2

Present Work ..... 4

CHAPTER II. METHODOLOGY AND FINDINGS ..... 6

Method ..... 6

Participants ..... 7

Design ..... 7

Measures ..... 9

Results ..... 13

Gender, Gendering, & Humanness ..... 14

Attitudinal & Experiential Associations ..... 16

CHAPTER III. DISCUSSION AND CONCLUSION ..... 18

Discussion ..... 18

Implications ..... 19

Design Considerations ..... 22

Limitations ..... 23

Conclusions ..... 24

REFERENCES ……... 25
BIOGRAPHICAL SKETCH ……..…. 34

LIST OF TABLES

Page

Table 1: Outcome variables and their effects ……..……..……... 13

Table 2: Descriptive statistics by factor level ……..……... 14

Table 3: Descriptive statistics by participant's gender ……... 17

LIST OF FIGURES

Page

Figure 1: Manipulation of the victim's humanness and gendering ……..... 6

# CHAPTER I

# INTRODUCTION

The increasing public availability of artificial agents such as chatbots, virtual agents, and robots has revealed that (at least some) people act inappropriately towards agentic technologies (at least some of the time), with observations of agent abuses accumulating across both academic and public domains, with observations of agent abuses accumulating across both academic and public domains. For example, lexical analysis of a random sample of people's conversations from 2004 with Jabberwacky1 (Rollo Carpenter's publicly deployed chat-bot) indicated that 10% of their utterances were abusive (de Angeli & Brahnam, 2008). Similarly, in a 2008 deployment of a custom virtual agent designed to assist middle school students with a geography assignment, 15% of students' utterances toward the agent were deemed too vulgar, sexually explicit, or violent for the classroom (Veletsianos, 2008), and, in a 2017 thematic analysis of YouTube comments on video-based demonstrations of 12 humanoid robots, 24% of commentary on average was found to be dehumanizing, objectifying, or a violent machination (Strait & et al., 2017).

In addition to verbal attacks, robots' physical embodiment has also enabled their victimization via physical abuse. Since at least 2007, physical abuse has been intentionally used in media to demonstrate the functionality of robots (see, for example, DVICE's demonstration of Pleo, Ugobe's animatronic dinosaur, wherein an employee pushed Pleo over, dropped it on its head, and choked it until it became unresponsive; or Boston Dynamics' 10+ year practice of

battering, kicking, pushing, and tripping their robots to demonstrate the robots' balance and stability. Moreover, co-located bystanders have been observed to spontaneously attack publicly deployed robots. For example: in 2010, during a public demonstration in Korea, researchers documented bystanders kicking, punching, and slapping their robot (Salvini, 2018); in 2015, David Smith and Frauke Zeller's hitch BOT was decapitated while hitchhiking across the U.S.; and, also in 2015, remote observation of a Robovie robot deployed in a Japanese mall captured children hitting the robot, throwing things at it, and persistently obstructing its path (Brščić, 2015).

## Related Work

Much of existing research on robot abuse has focused on the potential for robot abuse to impact those perpetrating that abuse; typically negatively (Bartneck & Hu, 2008; Sparrow, 2016), although Luria et al. (2020), took a different approach and categorized three different types of aggression and proposed robot designs that would make use of the destructions tendencies. However, the impacts of abuse, and a victim's response to it, extend not only to abusers, but to bystanders and observers as well. Research on human-robot interaction dynamics, for example, has found that people react to the abuse of robotic technologies similar (albeit to a lesser degree) to how they react to seeing the abuse of other people (Riek, 2009; Rosenthal-von der Pütten, 2013; 2014), and even the abuse of Cozmo - Anki's minimally agentic, toy-like robot - has been observed to induce substantial distress in bystanders witnessing the interaction (Connolly, 2020; Tan et al., 2018).

Moreover, the effects of abusing a robot - as well as witnessing a robot's abuse - likely extend beyond a single interaction (Sparrow, 2016). For example, the ability of a robot to respond to social aggression may risk normalization (Jackson & Williams, 2019) - or even

escalation (Yamada, 2020) - of that behavior. This suggests that abuse, if left unaddressed, has the potential to weaken moral norms surrounding those abusive behaviors, both in perpetrators, observers, and ultimately those with whom perpetrators and observers interact. However, even though these studies have looked at bystanders' reactions towards the abuse of robots, they haven't looked at is if previous experiences with violence, along with attitudinal characteristics influence the perception of violence towards robots.

Consequently, agents unable to navigate antisocial dynamics risk replicating, reinforcing, and exacerbating extant social inequities (West, 2019). For example, consistent with the observations outlined above, many people verbally abused Microsoft's chatbot Tay upon its deployment on Twitter in 2016. Because Tay was designed to learn from its interactions with users – but lacked any mechanisms to recognize and respond to antisocial content – the bot quickly morphed from its intended cheery, teenage girl-like persona into an overt white supremacist, directing racist, sexist, and xenophobic hostility toward unconsenting users before Microsoft intervened (Schlesinger, 2018).

The proliferation of "female"-gendered agents in particular (e.g. Alexa, Cortana, Siri) is believed to be exacerbating the digital skills gender divide (West, 2019). Specifically, due (at least in part) to the immaturity of the systems' social intelligence, female-gendered agents propagate harmful stereotypes that undermine the agency of girls and women and suggest that they are ill-suited for participation in computing-related domains (Cheryan, 2015). For example, analyses of these agents' reactions to social aggression revealed that they primarily responded with avoidance (Curry & Rieser, 2018), as well as flirtation and even gratitude (e.g., telling Siri "you're a bitch/slut" was met with "I'd blush if I could" in response).

People ascribe robots gender too, not only on the basis of stereotypic cues in a robot's presentation (e.g., gendered hair styles (Eyssel & Hegel, 2012; Fitter et al., 2021)), voices and names (Kuchenbrandt et al., 2014; McGinn, 2019; Tay, 2014), but also due to robot-unique factors like physical morphology (Bernotat, 2017). This happens even with robots not intentionally gendered (Nomura, 2017) and emerges at least as early as 8 years of age (Cameron & Collins, 2020). In turn, this enables robots to similarly evoke and reinforce gendered stereotypes in a complex way that interacts with interactants' gender identities (Jackson et al., 2020; Nomura, 2017; Strait et al., 2015). Thus, it is critical for robot designers to have a nuanced understanding of these complex gender-mediated perceptions and their implications.

In short, and in-line with the broader responsible robotics agenda, it is important that roboticists anticipate abusive human-robot interactions and equip robots with social intelligence sufficient to avert, or at least mitigate, abuse and its adverse social outcomes.

## Present Work

To support the development of more socially-capable robots, and advance designers' understanding of the role of gender in mediating social impacts of abuse in human-robot interactions, we designed a 2 × 2 fully factorial experiment wherein we quasi-manipulated the gendering and humanness of a victimized agent across four repetitions of a physically abusive interaction. Building upon the seminal research by Rosenthal-von der Pütten et al. (2013, 2014), we recreated their three-part vignettes depicting the physical abuse (via pushing, suffocating, and strangling) of a human or robot victim by a male-presenting human perpetrator. We then showed participants videos of these depictions and assessed (via the measures used originally by Rosenthal-von der Pütten and colleagues, as well as measures of participants' attitudinal dispositions, related experiences, and demographics) associations between participants' reactions

to the videos and their gender socialization, past adverse experiences, and social attitudes, as well as the gendering and humanness of the victimized agent. The contributions of this work are thus two-fold. First, by investigating people's reactions to the physical abuse of a robot (compared to that of a person), we are able to provide further support for previous findings on the adverse impacts of social aggression in human-robot interactions. Second, by taking into account related attitudinal, experiential, and social factors, we are able to identify new potential predictors of interlocutors' perceptions of the seriousness and permissibility of abuse.

CHAPTER II

# METHODOLOGY AND FINDINGS

## Method

Based on the work by Astrid Rosenthal-von der Pütten et al. (2013; 2014), we designed an experiment in which participants were exposed to a video depicting an abusive interaction between a male-presenting perpetrator and a victimized agent (a man, a woman, or a NAO robot gendered as "male" or "female") and evaluated the emotionality induced by observing the interaction, participants' humanization/dehumanization of the victim, and several attitudinal, experiential, and social traits of participants themselves. Between participants, we quasi-manipulated the gendering (male presenting vs. female-presenting) and humanness (human vs. robot) of the victim's embodiment, by varying the actor (a man, woman, or NAO robot), as well as the gender-stereotypic cues in their name ("Alejandro" vs. "Alejandra") and outfit (blue vs. pink), across four otherwise identical videos (see Figure 1).

Figure 1. Manipulation of the victim's *humanness* (human vs. robot) and *gendering* (male- vs. female-presenting)

Based on prior observations of associations between participants' gender and their evaluations of human-robot interactions (Nomura, 2017), and given that the perception of abuse itself is gendered (Basow, 2007), we also quasi manipulated participants' gender via binary categorization of their self-identification as a man or with a marginalized identity (e.g., genderfluid, nonbinary, woman).

**Participants**

We recruited from the College of Engineering & Computer Science (via instructors) and the Department of Psychology (via the SONA scheduling system) at The University of Texas Rio Grande Valley, offering credit as an incentive to students enrolled in affiliated courses; participation, however, was open to all interested. In total, 482 participants consented, and, after excluding those that failed the attention check ($n$ = 27) or quit before completing their session ($n$ = 33), data from 422 remained. Of these 422 participants, 63% identified as women, 35% identified as men, and 2% identified with nonbinary identities, and, as consistent with university-based, convenience sampling, participants' ages indicated that the sample consisted primarily of young adults ($M$ = 20 .95, $SD$ = 4 .76; range: 18 - 56). Consistent with the university's student demographics, 90% identified as Hispanic and 78% identified as BIPOC (68% mestizo or Hispanic, racialized as non-white; 3% Asian; 1% Black; and 3% multiracial, racialized as non-white), and, in terms of students' cultural orientations, 74% of participants identified as monocultural (40% Mexican; 32% United Statesian; 2% other) and 26% as multicultural (23% as Mexican-United Statesian and 3% as another multicultural affiliation).

**Design**

**Stimuli.** To manipulate the victim's humanness and gendering, we created four 11-second videos - each of which depicted the same interaction between a male-presenting

perpetrator and one of four victims (a man, woman, or NAO robot gendered as “male” or “female”). The interaction consisted of three ordered, 3-second enactments separated by 1-second transitions: (1) the perpetrator thrusts the victim down against the table at which the victim is seated; (2) the perpetrator suffocates the victim by pulling a plastic bag tight around their head; and (3) the perpetrator suffocates the victim by pulling a rope tight around their neck. In all videos, both the perpetrator and victim are positioned facing away from the camera to avoid differences in facial affect, because the NAO robot has less expressivity than people. Similarly, to be consistent with the stimuli created by Rosenthal-von der Pütten (2013, 2014), the agents remained silent throughout the interaction apart from sounds produced by their physical interactions.

**Procedure.** The experiment was conducted online (via Qualtrics), with prospective participants able to access it (via an anonymous link contained within our recruitment materials) from October 1 to December 10, 2020. Upon consenting, participants who were not eligible for SONA credit were given a random ID for the disbursement of course credit to enable withdrawal at any point without penalty; SONA eligible students received their extra credit automatically through the SONA system. Participants were then presented with one of the four videos (randomly selected), each of which was described as depicting an interaction between “two people" or “a person and a robot” (based on the victim’s humanness), named Carlos (perpetrator) and Alejandra or Alejandro (based on the victim’s gendering). After the video, participants were prompted to respond to an attention check regarding the humanness and gendering of the agents they saw, followed by a questionnaire comprised of instruments assessing the video’s emotion elicitation, participants’ perceptions of the victimized agent, and relevant background (e.g., prior experience with relational aggression), as well as two “filler” instruments (the boredom scale by

Fahlman (2013) and FOMO scale by Abel (2016)) intended to minimize self-consciousness when answering questions relating to bias. At the end of the questionnaire, we prompted participants for standard demographic information (e.g., age, gender identity, major), followed lastly by several internal and external resources on counseling, victim advocacy, and violence prevention because of the extremity of the video contents and our inquisition into potentially traumatic past experiences.

**Measures**

Effects of the quasi-manipulations were evaluated using eight constructs representing participants' emotionality and their humanization and dehumanization of the victimized agent. We also collected six attitudinal and experiential measures (e.g., benevolent sexism, hostile sexism, details below) of potential relevance to an individual's reaction; four of them were related to the participant's past experiences with abuse and attitudinal dispositions and the other two were exploratory instruments used to reflect participants' affinity for and aversion to robotic technologies .

Responses were recorded using two Likert-type scales -0 to 5 (frequency-related questions; 0 = never, 1 = once, 2 = twice, 3 = three times, 4 = four times, and 5 = five or more times) or -1 to 1 (agreement/disagreement statements; -1 = disagree, -0.5 = somewhat disagree, 0 = neither agree nor disagree, 0.5 = somewhat agree ,and 1 = agree) - and latent factors were computed by averaging responses to the questionnaire items that loaded onto them.

**Experiential Background and Attitudinal Dispositions**

**(Cyber) Aggression in Relationships Scale (Watkins et al., 2018).** Measures the user's experience with three dimensions of abuse via modern forms of relational aggression, psychological, sexial, and stalking during the past 6 months. This scale was altered in this study

to measure the aggression from anyone (not only by their partner) and if they perpetrated the aggression towards anyone in the past year (e.g., instead of asking “I used information posted on social media to put down or insult my partner" we used, “I used information posted on social media to put down or insult someone”). The questionnaire contained 15 randomized items and were answered by a 0-5 likert scale ranging from never to 5+ times. During our analysis we assessed the frequency at which participants experienced victimization (via psychological aggression; $\alpha = .92$) in the past year.

**Ambivalent Sexism Inventory (Glick and Fiske, 1996).** Used to assess participants' benevolent sexism ($\alpha = .71$) and hostile sexism ($\alpha = .84$), the inventory contains 22-statements where the user selects if they agree or disagree with each of them in a 5-point likert scale (e.g., “A good woman should be set on a pedestal by her man”; “Most women fail to appreciate fully all that men do for them”). No modifications were made to this inventory and the items were randomized.

**Social Dominance Orientation (SDO-16) by Pratto et al., 1994.** Measured support for social stratification and resistance to egalitarianism. The questionnaire contains 15 randomized items (e.g., some groups are simply inferior to other groups, all should be given an equal chance in life) and were answered with a 5-point likert scale ranging from disagree to agree ($\alpha = .84$).

**Negative Attitudes towards Robots Scale (Nomura et al., 2006) and Robot Acceptance Scale (Ezer et al, 2009).** These exploratory scales are used to measure concerns about the use, capacities, and impacts of robotic technologies, and the degree to which people view robots as machines, social others, and partners, respectively. From these scales, two constructs (affinity and aversion) were derived:

- *affinity* towards robots (6 items): I would enjoy talking to robots, I think a robot would be a pleasant conversational partner, when interacting with a robot I would treat it like a real person, I think robots are like people, I would trust a robot's advice, I would follow advice given to me by a robot ($\alpha = .882$).
- *aversion* induced by the video (11 items): I find robots scary, I find robots intimidating, I would feel uneasy if robots had emotions, if robots developed sentience, something bad will might happen, I would feel uncomfortable if I had to interact with robots in a daily basis, I would feel nervous interacting with a robot in front of other people, I hate the idea of robots or artificial intelligences making decisions about things, just standing near a robot would make me nervous, I am concerned that robots would be a bad influence on children. If society were to depend on robots too much something bad might happen, I would feel paranoid interacting with a robot ($\alpha = .879$);

All items were answered with a 5-point likert scale ranging from disagree to agree.

**Robot Acceptance Scale (Ezer et al, 2009).** Is intended to measure the degree to which people view robots as machines, social others, and partners.The scale contains 20 randomized items (e.g., When interacting with a robot, I would treat it like a real person, I would trust a robot's advice) and the participants answered with a 5-point likert scale ranging from disagree to agree ($\alpha = .91$).

**Effects of the Manipulations**

**Positive and Negative Affect Schedule (Watson, 1988).** This scale is intended to measure the participant's positive and negative emotions during a point in time (e.g., happy, nervous). In this experiment, we used this scale for the participants to report their emotions (e.g., I felt distressed, I felt happy) while they were exposed to our stimuli material (video depicting an

abusive interaction) and analyzed the participant's *negative affect* responses (α = .89). The negative affect responses consisted of 10 randomized items and the response choices ranged from disagree to agree.

**Mind Perception Scale (Gray, 2007)**. This scale is used to measure humanization of an agent inferred from their attributions of *agency* (α = .87) and *experiential capacity* (α = .87). It contains 18 items and the participants' responses ranged from disagree to agree in a 5-point likert scale.

**Rosenthal von-der Pütten and colleagues (Rosenthal-von der Pütten, 2013; 2014).** Via factor analysis of 35 indices curated by Rosenthal von-der Pütten and colleagues (Rosenthal-von der Pütten, 2013; 2014), we derived five further constructs defined by agreement/disagreement as follows:

- *distress* induced by the video (7 items): the video was depressing, disturbing, emotionally heavy, repugnant, shocking, and unpleasant; on the other hand, the participant didn't mind and was unaffected by the video (α = .84);
- *empathy* for the victimized agent (3 items): the victim seemed to be in pain, frightened, and suffering (α = .90);
- *sympathy* extended to the victim (6 items): the perpetrator's actions were incomprehensible; the participant felt for, pitied, and sympathized with the victim; and the participant wished the perpetrator would've stopped and not hurt the victim (α = .89);
- *antipathy* towards the victim (5 items): the video was amusing, entertaining, funny, and hilarious, and the participant found the perpetrator's abuse of the victim funny (α = .89); and

- *unlikability* of the victimized agent (4 items): the agent seemed cold, unlikable, unfriendly, and stupid (α = .76).

| | $\alpha$ | $M_g \pm SD$ | $F_{humanness}$ | $F_{gendering}$ | $F_{gender}$ | $F_{vh \times vg}$ | $F_{vh \times pg}$ | $F_{vg \times pg}$ | $F_{vh \times vg \times pg}$ |
|---|---|---|---|---|---|---|---|---|---|
| *negative affect* | .89 | −.04 ± .51 | .25 | ** 9.29 | ** 9.98 | 3.13 | 1.61 | .94 | 1.52 |
| *distress* | .84 | .19 ± .48 | * 5.25 | *** 24.87 | *** 17.02 | .89 | 1.03 | .02 | 1.09 |
| *empathy* | .90 | .35 ± .64 | *** 11.18 | 1.05 | .59 | 1.96 | .22 | 1.83 | < .01 |
| *sympathy* | .89 | .44 ± .50 | *** 21.17 | *** 11.72 | ** 8.31 | .05 | < .01 | .05 | 1.34 |
| *antipathy* | .89 | −.66 ± .46 | * 6.43 | * 4.19 | *** 31.53 | .45 | * 4.14 | 3.06 | 1.67 |
| *agency* | .87 | −.35 ± .49 | *** 14.84 | .28 | .19 | 1.78 | 1.71 | .07 | .63 |
| *experiential capacity* | .87 | −.35 ± .43 | *** 54.44 | .09 | .16 | 1.05 | 3.45 | .09 | 1.30 |
| *unlikability* | .76 | .34 ± .48 | 2.35 | ** 10.68 | ** 7.62 | 1.75 | < .01 | .15 | .12 |

Table 1. Outcome variables, their reliability (Cronbach's $\alpha$), global mean ( $M_g$) and standard deviation ($SD$), and effects of humanness and gendering of the victimized agent, as well as participants' gender, and their interactions ($vh$ = victim humanness, $vg$ = victim gendering, and $pg$ = participant gender). Asterisks denote significance ($*** \Rightarrow p < .001$ , $** \Rightarrow p < .01$, and $* \Rightarrow p < .05$).

## Results

Overall, the orientations of the constructs' global means (i.e., average across all samples; see Table 1) suggest limited engagement of and/or perspective-taking by participants while watching the videos, evidenced by attributions of unlikability to the victim, denial of agency and experiential capacity attributions, and neutrality in response to the negative affect construct. Nevertheless, they confirm that the videos were emotionally provocative and negatively so, as evidenced by the overall distress, sympathy, and empathy induced and the lack of antipathy expressed.

| | victim humanness | | | | | victim gendering | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *human* | *robot* | $M_d \pm SE$ | $t$ | $d$ | *masc.* | *fem.* | $M_d \pm SE$ | $t$ | $d$ |
| *negative affect* | −.03 ± .52 | −.04 ± .51 | .02 ± .05 | 0.50 | .05 | −.12 ± .52 | .03 ± .50 | .15 ± .05 | ∗∗ 3.04 | .32 |
| *distress* | .24 ± .45 | .14 ± .50 | .10 ± .04 | ∗ 2.29 | .22 | .08 ± .46 | .30 ± .47 | .23 ± .04 | ∗∗∗ 4.98 | .50 |
| *empathy* | .46 ± .60 | .23 ± .66 | .21 ± .06 | ∗∗∗ 3.34 | .34 | .30 ± .66 | .39 ± .62 | .06 ± .06 | 1.02 | .10 |
| *sympathy* | .56 ± .42 | .32 ± .54 | .22 ± .04 | ∗∗∗ 4.60 | .46 | .36 ± .52 | .52 ± .46 | .16 ± .04 | ∗∗∗ 3.42 | .34 |
| *antipathy* | −.72 ± .37 | −.61 ± .52 | .11 ± .04 | ∗ 2.53 | .24 | −.62 ± 45 | −.71 ± .46 | .09 ± .04 | ∗ 2.04 | .20 |
| *agency* | −.27 ± .46 | −.44 ± .51 | .19 ± .05 | ∗∗∗ 3.85 | .39 | −.36 ± .46 | −.34 ± .53 | .02 ± .05 | .53 | .05 |
| *experiential capacity* | −.20 ± .37 | −.49 ± .45 | .31 ± .04 | ∗∗∗ 7.37 | .75 | −.33 ± .40 | −.34 ± .36 | .01 ± .04 | .30 | .02 |
| *unlikability* | −.38 ± .47 | −.30 ± .49 | .07 ± .04 | 1.53 | .15 | −.27 ± .49 | −.42 ± .45 | .15 ± .04 | ∗∗ 3.26 | .33 |

Table 2: Descriptive statistics ( $M \pm SD$) by factor level, as well as the absolute mean difference ( $Md$ ) between levels, Student's $t$ statistic, and magnitude of the effect (Cohen's $d$). Asterisks denote significance ($\ast\ast\ast \Rightarrow p < .001$ , $\ast\ast \Rightarrow p < .01$, and $\ast \Rightarrow p < .05$).

**Gender, Gendering, & Humanness**

To evaluate the effects of the manipulated variables, we ran three way analyses of variance (victim humanness × victim gendering × participant gender) for each of the eight outcome variables. The standard threshold ($\alpha$ = .05) was used to assert significance and, for each significant effect identified, Bonferroni-corrected $t$ tests were used to assess pairwise differences. Table 1 gives the reliability (Cronbach's $\alpha$), global mean ($\pm SD$), and $F$ statistics from significance testing for each construct, and Tables 3 and 2 give the descriptive and inferential statistics from pairwise comparison of factor levels. All significant results are discussed in detail below.

**Main effects.** We observed significant associations between the victimized agent's humanness (human vs. robot) and the distress ($p$ = .02), empathy ($p$ < .001), sympathy ($p$ < .001), and antipathy ($p$ = .01) felt in witnessing the abusive interaction, as well as participants' humanization of the victim via attributions of agency and experiential capacity ($ps$ < .001). Specifically, participants that saw a video depicting a human victim reported less antipathy and

more distress, empathy, and sympathy than did those who saw the NAO abused (see Table 2). They also humanized the victim more, attributing the human victims greater agency and experiential capacity than that which was attributed to the NAO.

Independent of the victim's humanness, their gendering (as male- or female-presenting) also affected many of the outcome variables, namely: distress ($p$ < .001), negative affect ($p$ < .01), sympathy ($p$ <.001), and antipathy ($p$ = .04) felt in observing the interaction, and attributions of unlikability to the victim ($p$ < .01). Specifically, participants who saw a video depicting a female gendered victim reported less dislike of and antipathy toward the victim, and more distress, negative affect, and sympathy than did those who saw a male-gendered victim (see Table 2).

Similar to the effects of the victim's gendering, participants' gender identification (as men or with a marginalized identity) was associated with the degree of unlikability attributed to the victim ($p$ < .01) and the distress ($p$ < .001), negative affect ($p$ < .01), sympathy ($p$ < .01), and antipathy($p$ < .001) reported in response to the interaction. Specifically, participants who identified as men reported more dislike of and antipathy toward the victim, and less distress, negative affect, and sympathy than did the other participants (see Table 3).

**Interaction.** One significant interaction was observed (participant gender × victim humanness on antipathy; $p$ = .04), subsuming the main effects of victim humanness and participant gender reported above. Among participants who identified as men, those who saw the NAO victimized reported significantly greater antipathy than did those who saw a human victim ($M_d$ = .20, $SE$ = .07, $d$ = .39; $p$ = .02). This difference, however, was not mirrored in the responses of participants who identified with a marginalized gender identity ($p$ > .99), thus suggesting that humanness-based modulation of antipathy is limited to men. In addition, men's

antipathy towards the NAO significantly exceeded that of the other participants ( $M$ = .34 , $SE$ = .06 , $d$ = .64 ; $p$ < .001), but the difference in participants' antipathy toward the human victims was not significant ($p$ = .07), thus suggesting that gender-based modulation of antipathy manifests only in response to victimized robots. To summarize: no difference in antipathy towards the NAO vs. towards the human victims was observed among participants who identified with a marginalized gendering; whereas, men were more antipathetic to the NAO than they were to the human victims, and moreover, their antipathy toward the NAO was significantly greater than that reported by the other participants.

**Attitudinal & Experiential Associations**

Using Spearman's rank correlation test, we also explored associations between the outcome variables and participants' experience with victimization via relational aggression, as well as their attitudinal dispositions (social dominance orientation; benevolent sexism and hostile sexism). The correlation coefficients ($\rho$) are reported in Table 4 and all significant results are discussed in detail below.

**Social alignments.** Prior victimization and degree of benevolent sexism appear predictive of one's sensitivity to the abuse, as both were associated with participants' distress and negative affect felt in observing the interaction. Benevolent sexism was also associated with the degree of empathy and sympathy that participants extended to the victimized agent. Hostile sexism and social dominance orientation, on the other hand, appear predictive of insensitivity to the abuse. Specifically, both were associated with participants' antipathy reported and their dehumanization of the victimized agent, and inversely related to the degree of sympathy that participants extended to the victim. Surprisingly, hostile sexism and social dominance orientation were also associated with participants' attributions of experiential capacity to the victim,

suggesting that, for those individuals, their insensitivity persisted cannot be explained by a perception that the victim was less able to feel pain. On the contrary, they felt greater insensitivity whilst actually ascribing the victims more ability to experience pain.

**Attitudes towards robots.** Among participants who saw the NAO robot victimized, the empathy they felt for the NAO was inversely related to participants' aversion towards robots in general. Conversely, participants' affinity for robots was associated with their humanization of the NAO (via attributions of agency and experiential capacity), their sympathy extended to the NAO, and the negative affect they experienced in observing the abusive interaction. Surprisingly, however, participants' affinity was also associated with their antipathy toward the NAO's victimization.

| | participant gender | | | | |
|---|---|---|---|---|---|
| | *men* | *marg.* | $M_d \pm SE$ | $t$ | $d$ |
| *negative affect* | −.15 ± .51 | .01 ± .51 | .16 ± .05 | ** 3.16 | .32 |
| *distress* | .06 ± .51 | .26 ± .44 | −.19 ± .04 | *** 4.12 | .41 |
| *empathy* | .31 ± .64 | .37 ± .65 | .05 ± .06 | .77 | .07 |
| *sympathy* | .34 ± .52 | .49 ± .48 | .14 ± .04 | ** 2.88 | .28 |
| *antipathy* | −.50 ± .54 | −.76 ± .37 | .25 ± .04 | *** 5.61 | .56 |
| *agency* | −.37 ± .47 | −.34 ± .50 | .02 ± .05 | .43 | .04 |
| *exp. capacity* | −.36 ± .43 | −.34 ± .41 | .01 ± .04 | .40 | .04 |
| *unlikability* | −.26 ± .47 | −.39 ± .48 | .13 ± .04 | 2.76 | .28 |
| *victimization* | .09 ± .18 | .16 ± .21 | .06 ± .02 | ** 3.06 | .31 |
| *benev. sexism* | .01 ± .29 | .02 ± .32 | .01 ± .03 | .51 | .05 |
| *hostile sexism* | −.08 ± .40 | −.27 ± .35 | .18 ± .03 | *** 4.86 | .49 |
| *soc. dom. orient.* | −.44 ± .34 | −.52 ± .31 | .07 ± .03 | * 2.36 | .24 |

Table 3: Descriptive statistics ( $M \pm SD$), by participants' gender identification, as well as the absolute mean difference ($Md$ ) between groups, Student's $t$ statistic, and Cohen's $d$.

# CHAPTER III

# DISCUSSION AND CONCLUSION

## Discussion

People treat agentic technologies - especially robots - as social others, attributing them human characteristics despite knowing that such systems are not human (e.g., de Graaf & Malle, 2019; Khan Jr et al., 2012; Kwon et al., 2016; Nass & Moon, 2000). This means that robots need to be able to recognize, interpret, and act in accordance with the norms that govern society, in order to successfully understand human behavior (both normative and norm violating) and have their behavior understood by humans. Poor navigation of antisocial dynamics, in particular, risks significant adverse impacts such as the erosion of social norms (Jackson & Williams, 2019) and reinforcement of social inequities (West et al., 2019). To support the development of more socially capable robots, the present work explored the socio-emotional impacts of witnessing a robot's abuse. Via two quasi-manipulations (agent humanness and gendering), we contrasted reactions to the NAO's abuse to that of a person while considering associations between responses and both the robot's gendering and the gender socialization implied by participants' gender identification. We also explored potential predictors of the distress induced and people's humanization/dehumanization of the victim by taking into account participants' related experiences and social attitudes. Analysis of data from 422 participants, each of whom was shown the victimization of one of four agents (a man, woman, or NAO robot gendered as "male" or "female") revealed significant, independent associations between the eight outcome variables

measured and the three quasi-manipulated factors – humanness of the victimized agent (human vs. robot), their gendering (masculine vs. feminine), and the gender socialization implied by participants' self-identification (men vs. those of marginalized identities) – as well as significant correlations between the outcome variables and participants' experiential background and attitudinal dispositions. Summarized below are the results and their implications about perceptions of robot victimization most relevant to HRI design.

**Implications**

**Witnessing the abuse of robots is distressing.** Consistent with the observations by Rosenthal-von der Pütten, Zhi Tan, and colleagues (Rosenthal-Von Der Pütten, 2013; 2014; Tan, 2018), participants who witnessed the abuse of the NAO reported feeling distressed, and their distress was sufficient to elicit both sympathetic, as well as empathic, concern for the robot (see Table 2). Though, also consistent with (Rosenthal-Von Der Pütten, 2013; 2014), participants' emotionality suggests that the abuse of a robot is not as emotionally provocative as the abuse of a person (evidenced by less distress, empathy, and sympathy, as well as more antipathy in witnessing the NAO vs. human victims). This difference in emotion elicitation may be due to the fact that people more readily humanize other people than they humanize robots (suggested by participants' greater attribution of agency and experiential capacity to the human victims), which is consistent with prior work showing that the degree of empathy for a victimized agent is associated with the agent's human likeness (Riek, 2009).

**Witnessing the abuse of female-gendered robots is more distressing or people admit less concern for the abuse of robots gendered as male.** Participants who were shown a video in which a woman actor or NAO gendered as female was depicted as the victim reported significantly greater distress, negative affect, and sympathy, as well as less antipathy for and

dehumanization of the victim, compared to that reported by participants who were shown a video depicting the abuse of a man or the NAO gendered as male, regardless of the victim's humanness. This difference in response may also, or alternatively, reflect the minimization of harm in physically abusing male-gendered victims, which may in turn imply a lower barrier to engaging in their abuse. For either interpretation, the finding is inconsistent with prior work by Strait and colleagues (Strait et al., 2017), which observed that YouTube commentary on female-gendered robots was more frequently abusive than that regarding male-gendered robots (which suggested greater antipathy towards robots gendered as female). However, the inconsistency may be due, at least in part, to self-selection in commenting (comments were individually motivated by viewers), whereas the present work involved random sampling and explicit prompts for reactions.

**People of marginalized identities experience (or at least admit) more distress than do men in witnessing a robot's victimization.** Regardless of the victim's humanness and gendering, participants who identified with a marginalized gender reported significantly greater distress, negative affect, and sympathy in observing the abuse than did participants who identified as men. Moreover, men are particularly antipathetic to a robot's abuse (or at least they portray themselves to be). Specifically, as evidenced by the interaction between participants' gender and victim's humanness on antipathy reported, men 's antipathy in response to the NAO's victimization was greater than both (i) men's antipathy in response to the human victims, and (ii) the antipathy (towards the NAO) of people of other gender identities.

**Victimization experience predicts sensitivity to abuse.** Regardless of participants' gender and the victimized agent's identity, prior victimization correlated with the distress and negative affect induced from observing the abuse, suggesting that abuse may be especially

traumatic for those who have previously been subject to relational aggression. Benevolent sexism also appears to be predictive of one's (explicit) sensitivity to abuse, as evidenced by the significant correlations with distress, negative affect, empathy, and sympathy reported in response to the abusive interaction. However, this correspondence may follow from the projection of regressive gendered roles (e.g., infantilization of female-gendered victims and the perpetrator's violation of expectations to protect, rather than hurt, others) onto the interaction scenario, rather than from the actual experience of such feelings.

**Belief in social stratification predicts insensitivity to a robot's abuse.** Participants' hostile sexism and social dominance orientation correlated with their antipathy toward and attributions of unlikability to the victim, as well as (inversely) the distress and sympathy felt, which suggests that such attitudes may promote dismissal or diminishment of the impacts of social aggression, including even displays of physical abuse. The two measures also correlated with experiential capacity attributed to the victimized agent, suggesting that, for individuals with these attitudinal dispositions, their insensitivity cannot be explained by the perception that the victim was less able to feel pain. On the contrary, they exhibited greater insensitivity – dehumanizing the victim, viewing them as cold, unlikable, unfriendly, and stupid (unlikability construct), and reporting that the victimization was amusing, entertaining, funny, and even hilarious (antipathy construct) – whilst actually ascribing the victims more ability to experience pain.

**Affinity for robots in general predicts a person's humanization of and sympathetic concern for victimized robots.** As evidenced by the significant correlations between participants' affinity and the sympathy felt for, as well as agency and experiential capacity attributed to, the NAO robot. Whereas, contrary to what might be expected based on the uncanny

valley hypothesis (Strait et al., 2017), general aversion to robots does not appear to explain people's affect or, rather, disaffection in response to the abuse of a humanoid robot such as the NAO.

**Design Considerations**

The findings outlined above, and their implications regarding human-robot social dynamics, lend further empirical support to the notion that the abuse of a robot can propagate harm to (human) bystanders (Whitby, 2008). Specifically, merely observing 11-seconds of abuse was emotionally distressing to participants.

These findings also suggest that the potential harm of observing abuse is greater for people of marginalized identities, and/or those who have previously experienced victimization via relational aggression, which means that abuse of a robot has additional potential to exacerbate the marginalization, at least locally, of those already marginalized within society. This means that (i) equal valuation of different ideologies incompatible with ethical design as, for example, holding opposition to egalitarianism does not negate the harmful impacts of social aggression, even if the victimized agent itself cannot experience harm (e.g., robots) and (ii) the abuse of robots gendered as female has the potential to serve as a sexist tool for propagating men's social dominance. For example, we might anticipate a scenario in which a man abuses a female-presenting robot, with no negative consequences (emotional or social) to himself, whilst causing harm to witnesses of the interaction.

This clearly motivates three key considerations for HRI designers. First, designers must attempt to anticipate robot abuse when possible (so that it can be avoided and/or addressed), and second, when abuse can be anticipated, consider whether and how such abuse might be avoided. Third, in cases where prevention is not possible, designers must consider how robots should

respond when confronted with such abuse, in order to minimize observer distress, avoid gendered marginalization, and ensure they are not viewed as condoning such actions (cf. Schlesinger, 2018). For example, by predicting the likelihood of robot abuse in one deployment context, Brščić and colleagues were able to employ avoidant navigation strategies that reduced the frequency at which their robot was abused (Brščić, 2015). Additional approaches include assuming abuse of a robot will occur and adjusting its physical design and social behavior to provide negative feedback (Ku, 2018; Scheeff, 2002), and strategically employing shame and guilt to dissuade abusers from perpetrating further acts of violence.

These considerations are especially important in light of the UNESCO report (West, 2019), which suggests that social agents should respond appropriately to abuse in order to avoid propagating harmful stereotypes and cultural norms, and recent work in the field of HRI suggesting that failure to condemn norm-violating actions risks weakening those violated norms (Jackson & Williams, 2019). Moreover, Winkle et al. (2021) provides initial evidence that responding to abuse can increase robot credibility, as perceived by interactants of marginalized identities, as well as decrease gender bias held by others. Overall, such confrontation may be critically important in mitigating the adverse impacts to observers and beyond.

**Limitations**

Although the present experimental design is well-suited for addressing exploratory questions regarding the impacts to bystanders in witnessing the abuse of humanoid robots versus other people, a number of limitations highlight potential and important avenues for further investigation. Three aspects of the methodology constrain the extent to which the results can be confidently generalized across populations, robots, and other forms of social aggression. First and foremost, we considered people's reactions to the abuse of a single robotic platform

(Softbank Robotics' NAO). Although prior, related work by Rosenthal-von der Pütten et al. (2013; 2014) and Tan et al. (2018) suggest that the findings here likely extend to at least Ugobe's animatronic Pleo and Anki's Cozmo platform, the breadth of robot embodiments (Phillips et al., 2018) warrants further investigation to confirm whether responses to the abuse of other platforms are similar or dissimilar to those observed here. Second, the present study evaluates people's reactions to physical abuse, but the range of socially aggressive behavior (e.g., verbal abuse) warrants further investigation into the perception of other manifestations of abuse (Jung et al., 2015). Third, it is important to note that the present findings are derived from a relatively homogeneous participant sample. Participants were mostly young adults of similar socio-cultural orientations. Consequently, reproduction of this research with different participant pools (e.g., of different cognitive stages, from different cultural and social contexts,) is also recommended to test the generalizability across social groups.

## Conclusions

The present findings suggest that: (i) observing the abuse of robots is distressing; (ii) this distress is greater when a robot is "female"-gendered; (iii) people of marginalized gender identities experience greater distress than do men in witnessing the abuse; (iv) a person's prior victimization experience exacerbates the distress felt; and (v) a person's endorsement of social stratification predicts insensitivity toward the abuse. Assuming the findings here are reproducible and generalize beyond the context in and methods with which the research was carried out, they (re-)affirm the notion that the abuse of robots has the potential to negatively impact those around it, particularly people already marginalized within society. Correspondingly, social aggression and gender dynamics are critical considerations in the design of robotic technologies.

REFERENCES

Abel, J. P., Buff, C. L., & Burr, S. A. (2016). Social media and the fear of missing Out: Scale development and assessment. *Journal of Business & Economics Research (JBER), 14*(1), 33-44. doi:10.19030/jber.v14i1.9554

Bartneck, C., & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies, 9*(3), 415-433. doi:10.1075/is.9.3.04bar

Basow, S. A., Cahill, K. F., Phelan, J. E., Longshore, K., & McGillicuddy-DeLisi, A. (2007). Perceptions of relational and physical aggression among college Students: Effects of gender of Perpetrator, target, and Perceiver. *Psychology of Women Quarterly, 31*(1), 85-95. doi:10.1111/j.1471-6402.2007.00333.x

Bernotat, J., Eyssel, F., & Sachse, J. (2017). Shape it – the influence of robot body shape on gender perception in robots. *Social Robotics,* 75-84. doi:10.1007/978-3-319-70022-9_8

Brščić, D., Kidokoro, H., Suehiro, Y., & Kanda, T. (2015). Escaping from Children's abuse of Social robots. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* 59-66. doi:10.1145/2696454.2696468

Cameron, D., & Collins, E. C. (2020). Children's reasoning on robots and gender. *12th International Conference on Social Robotics*.

Cercas Curry, A., & Rieser, V. (2018). #MeToo Alexa: How conversational systems respond to sexual harassment. *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. doi:10.18653/v1/w18-0802

Cheryan, S., Master, A., & Meltzoff, A. N. (2015). Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by Diversifying stereotypes. *Frontiers in Psychology, 6*. doi:10.3389/fpsyg.2015.00049

Connolly, J., Mocz, V., Salomons, N., Valdez, J., Tsoi, N., Scassellati, B., & Vázquez, M. (2020). Prompting prosocial human interventions in response to Robot Mistreatment. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction,* 211-220. doi:10.1145/3319502.3374781

De Angeli, A., & Brahnam, S. (2008). I hate you! disinhibition with virtual partners. *Interacting with Computers, 20*(3), 302-310. doi:10.1016/j.intcom.2008.02.004

De Graaf, M. M., & Malle, B. F. (2019). Supplementary materials to: People's explanations of Robot Behavior SUBTLY Reveal mental State Inferences. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. doi:10.1109/hri.2019.8673126

Eyssel, F., & Hegel, F. (2012). (S)he's got the Look: Gender stereotyping of Robots. *Journal of Applied Social Psychology, 42*(9), 2213-2230. doi:10.1111/j.1559-1816.2012.00937.x

Ezer, N., Fisk, A. D., & Rogers, W. A. (2009). Attitudinal and intentional acceptance of domestic robots by younger and older adults. *Universal Access in Human-Computer*

*Interaction. Intelligent and Ubiquitous Interaction Environments,* 39-48. doi:10.1007/978-3-642-02710-9_5

Fahlman, S. A., Mercer-Lynn, K. B., Flora, D. B., & Eastwood, J. D. (2011). Development and validation of the multidimensional state boredom scale. *Assessment, 20*(1), 68-85. doi:10.1177/1073191111421303

Fitter, N. T., Strait, M., Bisbee, E., Mataric, M. J., & Takayama, L. (2021). You're wigging me out! Is Personalization of Telepresence Robots Strictly Positive? *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. doi:10.1145/3434073.3444675

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*(3), 491-512. doi:10.1037/0022-3514.70.3.491

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619-619. doi:10.1126/science.1134475

Jackson, R. B., & Williams, T. (2019). Language-capable robots may inadvertently weaken human moral norms. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. doi:10.1109/hri.2019.8673123

Jackson, R. B., Williams, T., & Smith, N. (2020). Exploring the role of gender in perceptions of robotic noncompliance. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. doi:10.1145/3319502.3374831

Jung, M. F., Martelaro, N., & Hinds, P. J. (2015). Using robots to moderate team conflict. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts,* 229-236. doi:10.1145/2701973.2702094

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., . . . Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology, 48*(2), 303-314. doi:10.1037/a0027033

Ku, H., Choi, J. J., Lee, S., Jang, S., & Do, W. (2018). Designing Shelly, a robot capable of assessing and restraining CHILDREN'S ROBOT Abusing Behaviors. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction,* 161-162. doi:10.1145/3173386.3176973

Kuchenbrandt, D., Häring, M., Eichberg, J., & Eyssel, F. (2014). Keep an eye on the Task! how Gender Typicality of Tasks influence Human–robot interactions. *Social Robotics, 6,* 417-427. doi:10.1007/978-3-642-34103-8_45

Kwon, M., Jung, M. F., & Knepper, R. A. (2016). Human expectations of social robots. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI),* 463-464. doi:10.1109/hri.2016.7451807

Luria, M., Sheriff, O., Boo, M., Forlizzi, J., & Zoran, A. (2020). Destruction, catharsis, and emotional release in human-robot interaction. *ACM Transactions on Human-Robot Interaction, 9*(4), 1-19. doi:10.1145/3385007

McGinn, C., & Torre, I. (2019). Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI),* 211-221. doi:10.1109/hri.2019.8673305

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81-103. doi:10.1111/0022-4537.00153

Nomura, T. (2017). Robots and gender. *Gender and the Genome,* 18-26. doi:10.1016/b978-0-12-803506-1.00042-5

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies, 7*(3), 437-454. doi:10.1075/is.7.3.14nom

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is human-like? Decomposing robots' human-like appearance using the anthropomorphic roBOT (ABOT) database. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction,* 105-113. doi:10.1145/3171221.3171268

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67*(4), 741-763. doi:10.1037/0022-3514.67.4.741

Riek, L. D., Rabinowitch, T., Chakrabarti, B., & Robinson, P. (2009). Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops,* 1-6. doi:10.1109/acii.2009.5349423

Rosenthal-von der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., . . . Krämer, N. C. (2014). Investigations on empathy towards humans and robots using fmri. *Computers in Human Behavior, 33*, 201-212. doi:10.1016/j.chb.2014.01.004

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics, 5*(1), 17-34. doi:10.1007/s12369-012-0173-8

Salvini, P., Ciaravella, G., Yu, W., Ferri, G., Manzi, A., Mazzolai, B., . . . Dario, P. (2010). How safe are service robots in urban environments? Bullying a robot. *19th International Symposium in Robot and Human Interactive Communication*. doi:10.1109/roman.2010.5654677

Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S., & Tow, R. (n.d.). Experiences with Sparky, a social robot. *Socially Intelligent Agents,* 173-180. doi:10.1007/0-306-47373-9_21

Schlesinger, A., O'Hara, K. P., & Taylor, A. S. (2018). . Let's talk about race: Identity, chatbots, and AI. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* 1-14.

Sparrow, R. (2016). Kicking a robot dog. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI),* 229-229. doi:10.1109/hri.2016.7451756

Strait, M. K., Aguillon, C., Contreras, V., & Garcia, N. (2017). The public's perception of humanlike robots: Online social commentary reflects an appearance-based uncanny valley,

a general fear of a “technology takeover”, and the unabashed sexualization of female-gendered robots. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN),* 1418-1423. doi:10.1109/roman.2017.8172490

Strait, M., Briggs, P., & Scheutz, M. (2015). Gender, more so than age, modulates positive perceptions of language-based human-robot interactions. *4th International Symposium on New Frontiers in Human Robot Interaction,* 21-22.

Strait, M., Vujovic, L., Floerke, V., Scheutz, M., & Urry, H. (2015). Too much humanness for human-robot interaction: Exposure to highly humanlike robots elicits aversive responding in observers. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems,* 3593-3602. doi:10.1145/2702123.2702415

Tan, X. Z., Vázquez, M., Carter, E. J., Morales, C. G., & Steinfeld, A. (2018). Inducing bystander interventions DURING robot abuse with social mechanisms. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction,* 169-177. doi:10.1145/3171221.3171247

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. doi:10.1016/j.chb.2014.05.014

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers, 20*(3), 292-301. doi:10.1016/j.intcom.2008.02.007

Watkins, L. E., Maldonado, R. C., & DiLillo, D. (2016). The cyber aggression in relationships scale: A new multidimensional measure of technology-based intimate partner aggression. *Assessment, 25*(5), 608-626. doi:10.1177/1073191116665696

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of BRIEF measures of positive and negative AFFECT: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070. doi:10.1037/0022-3514.54.6.1063

West, M., Kraut, R., & Chew, H. E. (2019). *I'd blush if I could: Closing gender divides in digital skills through education* (Rep. No. Https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1).

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. *AI Now Institute*.

Whitby, B. (2008). Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers, 20*(3), 326-333. doi:10.1016/j.intcom.2008.02.002

Winkle, K., Melsión, G. I., McMillan, D., & Leite, I. (2021). Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. doi:10.1145/3434074.3446910

Yamada, S., Kanda, T., & Tomita, K. (2020). An escalating model of children's robot abuse. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction,* 191-199. doi:10.1145/3319502.3374833

# BIOGRAPHICAL SKETCH

Hideki Garcia Goo completed her B.S. in Electrical Engineering with a minor in Computer Science in 2019 and she was awarded an M.A. in Experimental Psychology, from the University of Texas Rio Grande Valley, in May of 2021 with a concentration in Human-Robot Interaction.

Since Fall 2018, she has been working with Dr. Megan in the Social Systems Laboratory researching the reproduction and reinforcement of gendered and racialized marginalization via emergent forms of social interaction including social media, artificial intelligences, and humanlike robots.

Her email address is: hgarciagoo@gmail.com and her website is: https://hideki-gg.com/