Victims and Observers: How Gender, Victimization Experience, and Biases Shape Perceptions of Robot Abuse

Hideki Garcia Goo*, Katie Winkle, Tom Williams, and Megan K. Strait

***Abstract*— With the deployment of robots in public realms, researchers are seeing more and more cases of abusive disinhibition towards robots. Because robots embody gendered identities, poor navigation of antisocial dynamics may reinforce or exacerbate gender-based violence. Robots deployed in social settings must recognize and respond to abuse in a way that minimizes ethical risk. This will require designers to first understand the risk posed by abuse of robots, and how humans perceive robot-directed abuse. To that end, we conducted an exploratory study of reactions to a physically abusive interaction between a human perpetrator and a victimized agent. 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 (via use of a *human* actor vs. NAO *robot*) and gendering (via inclusion of stereotypically *masculine* vs. *feminine* cues in their presentation) across four video-recorded reproductions of the interaction. Analysis of data from 417 participants, each of whom watched one of the four videos, indicates that the intensity of emotional distress felt by an observer is associated with their gender identification, previous experience with victimization, hostile sexism, and support for social stratification, as well as the victim's gendering.**

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

I. Introduction

The increasingly public availability of artificial agents such as chatbots, virtual agents, and robots has revealed a tendency for some people to act inappropriately towards agentic technologies, with observations of abuse directed at artificial social agents accumulating across both academic and public domains [1], [2], [3], [4]. For physically embodied robots, these problems are only exacerbated, since in addition to verbal attacks, robots' physical embodiment has also enabled their victimization via physical abuse. For the purpose of this study, we will define robot abuse as the act of exhibiting aggressive behaviors towards them [5].

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). While these behaviors might not be intended to harm the robots, they can still be perceived as abusive by bystanders [6]. Co-located bystanders have also been observed to spontaneously attack publicly deployed robots. For example, during a public demonstration in Korea in 2010, researchers documented bystanders kicking, punching, and slapping their robot [7]; in 2015, David Smith and Frauke Zeller's *hitchBOT* was *decapitated* while hitchhiking across the U.S.; and in 2015, remote observation of a Robovie2 deployed in a Japanese mall captured children hitting the robot, throwing things at it, and persistently obstructing its path [8]. These observations have led to a variety of recent work seeking to understand the nature, extent, antecedents, and consequences of such abuse.

The present work supports 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 exploratory experiment wherein we quasi-manipulated the gendering and humanness of a victimized agent across four repetitions of a physically abusive interaction.

Building on the seminal research by Rosenthal-von der Pütten and colleagues [9], [10], 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 measures used originally by Rosenthal-von der Pütten and colleagues, as well as measures of participants' attitudinal dispositions,

*Hideki Garcia Goo is with Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twenteh.garciagoo@utwente.nl. Tom Williams with the Department of Computer Science, Mines University, twilliams@mines.edu.Katie Winkle with the Department of Information Technology, Uppsala University, katie.winkle@it.uu.se. Megan Strait performed this work with the University of Texas Rio Grande Valley

related experiences, and demographics) associations between participants' reactions to the videos and their gender socialization, past adverse experiences, social attitudes, and their 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 can 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 can identify new predictors of interlocutors' perceptions of the seriousness and permissibility of abuse.

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 [11], [12], although cf. [13]). However, the impacts of abuse, and a victim's response to it, extend not only to abusers, but to bystanders and observers as well. Although people may believe themselves disaffected in aggressing an artificial agent, **the impacts of abuse, and an agent's response to it, extend beyond individual interactions.** Research on human-robot interaction dynamics, for example, has found that people react to the abuse of robotic technologies similar (to a lesser degree) to how they react to seeing the abuse of other people [14], [9], [10], 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 [15], [5].

Moreover, the effects of abusing a robot – as well as witnessing a robot's abuse – likely extend beyond a single interaction (e.g., [12]). For example, the inability of a robot to respond to social aggression may risk normalization [16] – or even escalation [17] – 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.

Consequently, agents unable to navigate antisocial dynamics risk replicating, reinforcing, and exacerbating extant social inequities [18]. For example, consistent with the observations outlined above, many people verbally abused Microsoft's chatbot *Tay* upon its 2016 deployment. 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 [19].

People ascribe robots gender [20], even in the absence of intentional gender cueing [21], a phenomena that emerges at least as early as 8 years of age [22]. This enables robots to similarly evoke and reinforce gendered stereotypes in a complex way that interacts with interactants' gender identities [23], [21], [24]; but also allows for the intentional subversion of gender norms and stereotypes [25], [26]. Given differences in perceptions (and realities) of gender-based violence [27], [28], [29], [30], it is critical for robot designers to have a nuanced understanding of these complex gender-mediated perceptions and their implications in the context of robot abuse.

II. Method

Based on work by Astrid Rosenthal-von der Pütten and colleagues [9], [10], we conducted a between-subject study in which participants watched 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"). We evaluated the emotionality induced by observing the interaction, participants' humanization/dehumanization of the victim, and several attitudinal, experiential, and social traits of participants themselves. To manipulate the **gendering** (*male-presenting* vs. *female-presenting*) of the victim's embodiment, we used male and female actor actors, with gender cues in their names ("Alejandro" vs. "Alejandra") and outfits (blue vs. pink) (see Figure 1). All study materials and procedures were approved by the Institutional Review Board (IRB) at the University of Texas Rio Grande Valley under protocol IRB-20-0379.

Based on prior observations of associations between participants' gender and their evaluations of human-robot interactions (e.g., [21]), and given that the perception of abuse itself is gendered (e.g., [27]), we also categorized participants' **gender** via binary categorization of their self-identification as a *man* or with a *marginalized* identity (e.g., genderfluid, nonbinary, woman).

A. Participants

We recruited from the College of Engineering & Computer Science (via instructors) and from the Department of Psychology (via SONA system) at the University of Texas Rio Grande Valley (UTRGV), offering credit as an incentive. 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 417 remained; 63.55% identified as women, 34.77% identified as men, and 1.68% identified as transgender (for the purpose of this research we refer as transgender to people who transition from one gender to another [31]). The sample consisted primarily of young adults ($M = 20.93$, $SD = 4.75$; *range*: $18 - 56$). Consistent with the university's demographics, 89.93% identified as Hispanic (68.35% non-white; 18.70% white; $.48\%$ Asian; and 2.4% multiracial), 2.4% as Asian, $.48\%$ as Black, 3.60% as White, $.48\%$ as Other, and 3.11% preferred not to answer. In terms of students cultural orientations, 74% of participants identified as monocultural (40% Mexican; 32% United Statesian; and 2% other cultural orientations) and 26% as multicultural (23% as bicultural (Mexican-Statesian), and 3% as multicultural).

B. Design

Stimuli: We made four 11-second videos showing a male-presenting perpetrator interacting violently with a victim (man, woman, or NAO robot).The interaction consisted of

three ordered, 3-second enactments separated by 1-second transitions of the perpetrator: (1) thrusting the victim against the table; (2) suffocating the victim with a plastic bag; and (3) suffocating the victim with a rope. In all videos, both the perpetrator and victim are positioned facing away from the camera and the agents remained silent apart from sounds produced by their physical interactions.

Procedure: The experiment was conducted online (via Qualtrics), with prospective participants able to access it from October 1 to December 10, 2020. Participants were then randomly presented with one of the four videos, each of which was described as depicting an interaction between "two people" or "a person and a robot", named *Carlos* (perpetrator) and *Alejandra* or *Alejandro* (victim). 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 assessing the video's emotion elicitation, participants' perceptions of the victimized agent, and relevant background, as well as two "filler" instruments (the boredom and FOMO scales [32], [33]). The ordering of instruments and questions contained within each was randomized. At the end of the questionnaire, we prompted participants for standard demographic information, and provided internal and external resources on counseling, victim advocacy, and violence prevention. The duration of the survey was expected to be 40 minutes.

C. Measures

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: Using the (Cyber) Aggression in Relationships Scale [34], we assessed the frequency at which participants experienced *victimization* (via psychological aggression) in the past year, and, using the Ambivalent Sexism Inventory [35] and SDO-16 [36], we assessed participants' *benevolent sexism* (covert infantilization of women), *hostile sexism* (overt hostility toward women), and *social dominance orientation* (support for social stratification and resistance to egalitarianism). In addition, we derived two constructs reflecting *affinity* for and *aversion* to robotic technologies, using two exploratory instruments: (1) the Negative Attitudes towards Robots Scale ([37], intended to measure concerns about the use, capacities, and impacts of robotic technologies); and (2) the Robot Acceptance Scale ([38], to measure the degree to which people view robots as machines, social others, and partners).

Effects of the manipulations: Using the Positive and Negative Affect Schedule [39] and the Mind Perception Scale [40], we assessed participants' **negative affect** and humanization of the victimized agent (inferred from their attributions of **agency** and **experiential capacity**), and, via factor analysis of 35 indices curated by Rosenthal von-der Pütten and colleagues ([9], [10]), we derived five further constructs defined by agreement/disagreement as follows:

- **distress** induced by the video (8 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;
- **empathy** for the victimized agent (3 items): the victim seemed to be *in pain*, *frightened*, and *suffering*;
- **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;
- **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*; and
- **unlikability** of the victimized agent (4 items): the agent seemed *cold*, *unlikable*, *unfriendly*, and *stupid*.

III. Results

Overall, the orientations of the constructs' global means (i.e., average across all samples; see Table I) 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.

A. Gender, Gendering, & Humanness

To evaluate the effects of the manipulated variables, we ran three-way analyses of variance (*victim humanness* $\times$ *victim gendering* $\times$ *participant gender*) for each of the eight outcome variables. 69 participants who identified with a marginalized gender identity observed the video portraying the abuse of a male-presenting robot, 61 observed the abuse of a female-presenting robot, 70 observed the abuse of a male human, and 71 observed the abuse of a female human. For the participants who identified as male, 33 observed the abuse of a male robot, 45 observed the abuse of a female robot, 36 observed the abuse of a human male, and 31 observed the abuse of a female human.

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 I gives the reliability (Cronbach's α), global mean ($\pm$ SD), and F statistics from significance testing for each construct, and Tables III and II 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

	α	$M_g \pm SD$	$F_{humanness}$	$F_{gendering}$	F_{gender}	$F_{vh \times vg}$	$F_{vh \times pg}$	$F_{vg \times pg}$	$F_{vh \times vg \times pg}$
negative affect	.89	$-.05 \pm .51$	.09	** 7.33	*** 11.79	2.37	.76	1.24	.96
distress	.86	$.19 \pm .48$	* 5.54	*** 23.72	*** 16.59	.72	.98	.02	1.02
empathy	.90	$.35 \pm .65$	** 10.67	.56	.74	2.50	.41	2.20	.06
sympathy	.89	$.44 \pm .50$	*** 22.47	** 10.78	** 7.87	.02	< .01	.06	1.32
antipathy	.89	$-.68 \pm .45$	** 7.61	* 4.77	*** 27.77	.72	* 5.22	2.56	1.94
agency	.87	$-.36 \pm .50$	*** 13.84	.34	.37	1.86	1.48	.10	.43
experiential capacity	.88	$-.35 \pm .44$	*** 53.18	.07	.29	1.18	3.26	.13	.97
unlikability	.76	$-.34 \pm .48$	2.62	*** 11.64	* 6.68	1.49	.05	.15	.17

TABLE I

OUTCOME VARIABLES, THEIR RELIABILITY (CRONBACH'S α), 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 * $\Rightarrow p < .05$).

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 II). They also humanized the victim more, attributing greater agency and experiential capacity to human victims than 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* ($p < .001$), *sympathy* ($p < .01$), and *antipathy* ($p < .001$) felt in observing the interaction, and attributions of *unlikability* to the victim ($p < .001$). 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 II).

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 III).

Interactions: One significant interaction was observed (*participant gender* $\times$ *victim humanness* on *antipathy*; $p = .02$), 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 ($M_d = .22$, $SE = .07$, $d = .39$; $p = .02$). 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.

In short, participants who identified with a marginalized gender showed no difference in antipathy towards human and NAO victims. However, men showed more antipathy towards NAO victims than human victims, and their antipathy was greater than other participants.

B. 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 (ρ) are reported in Table IV 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. In summary, *hostile* sexism and *social dominance orientation* predict insensitivity to abuse and were associated with antipathy, dehumanization, and reduced sympathy towards the victim. Surprisingly, these factors were also linked to attributions of greater experiential capacity to the victim, indicating increased insensitivity despite acknowledging the victim's ability to feel pain.

Attitudes towards robots: Among participants who saw the NAO victimized, the empathy they felt for the NAO was inversely related to participants' *aversion* towards robots in general. 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

	victim humanness					victim gendering				
	human	*robot*	$M_d \pm SE$	t	d	*masc.*	*fem.*	$M_d \pm SE$	t	d
negative affect	$-.03 \pm .52$	$-.04 \pm .51$	$.01 \pm .05$	0.29	.03	$-.12 \pm .52$	$.03 \pm .50$	$.14 \pm .05$	$**$ 2.70	.28
distress	$.24 \pm .45$	$.14 \pm .50$	$.11 \pm .05$	$*$ 2.35	.23	$.08 \pm .46$	$.30 \pm .47$	$.23 \pm .05$	$***$ 4.82	.49
empathy	$.46 \pm .60$	$.23 \pm .66$	$.22 \pm .06$	$***$ 3.27	.34	$.30 \pm .66$	$.39 \pm .62$	$.05 \pm .07$	.75	.08
sympathy	$.56 \pm .42$	$.32 \pm .54$	$.23 \pm .05$	$***$ 4.74	.48	$.36 \pm .52$	$.52 \pm .46$	$.16 \pm .05$	$***$ 3.28	.33
antipathy	$-.72 \pm .37$	$-.61 \pm .52$	$-.12 \pm .04$	$**-2.76$	$-.27$	$-.62 \pm 45$	$-.71 \pm .46$	$-.09 \pm .04$	$*-2.18$	$-.22$
agency	$-.27 \pm .46$	$-.44 \pm .51$	$.19 \pm .05$	$***$ 3.72	.39	$-.36 \pm .46$	$-.34 \pm .53$	$.03 \pm .05$	.58	.06
exper. capacity	$-.20 \pm .37$	$-.49 \pm .45$	$.31 \pm .04$	$***$ 7.29	.76	$-.33 \pm .40$	$-.34 \pm .36$	$-.01 \pm .04$	$-.27$	.03
unlikability	$-.38 \pm .47$	$-.30 \pm .49$	$-.08 \pm .05$	-1.62	$-.16$	$-.27 \pm .49$	$-.42 \pm .45$	$-.17 \pm .05$	$***-3.41$	$-.35$

TABLE II

DESCRIPTIVE STATISTICS ($M \pm SD$) BY FACTOR LEVEL, AS WELL AS THE ABSOLUTE *mean difference* (M_d) BETWEEN LEVELS, STUDENT'S t STATISTIC, AND MAGNITUDE OF THE EFFECT (COHEN'S d). ASTERISKS DENOTE SIGNIFICANCE ($*** \Rightarrow p < .001$, $** \Rightarrow p < .01$, AND $* \Rightarrow p < .05$).

the abuse. Surprisingly, participants' affinity was also associated with their antipathy toward the NAO's victimization.

IV. DISCUSSION

People treat agentic technologies – especially robots – as social others, attributing them human characteristics despite being fully aware that such systems are not human (e.g., [41], [42], [43], [44]). This means that robots need to be able to recognize, interpret, and act in accordance with social norms 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 the erosion of social norms [16] and reinforcement of social inequities [18].

To understand the risks posed by abusive human-robot interactions, 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.

	participant gender				
	men	*marg.*	$M_d \pm SE$	t	d
neg. affect	$-.15 \pm .51$	$.01 \pm .51$	$-.18 \pm .05$	-3.43^{***}	$-.35$
distress	$.06 \pm .51$	$.26 \pm .44$	$-.19 \pm .05$	-4.07^{***}	$-.41$
empathy	$.31 \pm .64$	$.37 \pm .65$	$-.06 \pm .07$	$-.86$	$-.09$
sympathy	$.34 \pm .52$	$.49 \pm .48$	$-.14 \pm .05$	-2.80^{**}	$-.28$
antipathy	$-.50 \pm .54$	$-.76 \pm .37$	$.24 \pm .04$	5.27^{***}	.53
agency	$-.37 \pm .47$	$-.34 \pm .50$	$-.03 \pm .05$	$-.61$	$-.06$
exp. capacity	$-.36 \pm .43$	$-.34 \pm .41$	$-.01 \pm .04$	$-.54$	$-.05$
unlikability	$-.26 \pm .47$	$-.39 \pm .48$	$.13 \pm .05$	2.58^{*}	.26
victimization	$.09 \pm .18$	$.16 \pm .21$	$-.06 \pm .02$	3.01^{**}	$-.31$
benev. sexism	$.01 \pm .29$	$.02 \pm .32$	$-.01 \pm .03$	$-.43$	$-.04$
host. sexism	$-.08 \pm .40$	$-.27 \pm .35$	$.18 \pm .04$	4.77^{***}	.49
soc. dom. or.	$-.44 \pm .34$	$-.52 \pm .31$	$.07 \pm .03$	1.97^{*}	.20

TABLE III

DESCRIPTIVE STATISTICS ($M \pm SD$), BY PARTICIPANTS' GENDER IDENTIFICATION, AS WELL AS THE ABSOLUTE *mean difference* (M_d) BETWEEN GROUPS, STUDENT'S t STATISTIC, AND COHEN'S d.

Analysis of data from 417 participants, each of whom was shown the victimization of one of four agents (a man, woman, or NAO 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. Below we summarize the results and their implications most relevant to HRI design.

Witnessing the abuse of robots is distressing. Consistent with the observations by Rosenthal-von der Pütten [9], [10], and by Tan [5], participants who witnessed the abuse of the NAO reported feeling distressed, and their distress was sufficient to elicit both sympathetic, as well as empathetic, concern for the robot (see Table II). However, also consistent with [9], [10], the abuse of a robot was 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 may be because people tend to humanize humans more than robots, as shown by prior research, and empathy is associated with the agent's human likeness [14].

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 of those shown a video depicting the abuse of a man or the NAO gendered as male. This difference in response

	victimization	*benevolent sexism*	*hostile sexism*	*soc. dom. orientation*	*robot affinity*	*robot aversion*
negative affect	** .16	*** .17	.02	.00	** .01	.01
distress	** .15	** .14	−.06	** − .14	.13	.05
empathy	.01	*** .18	.08	.00	.11	* − .16
sympathy	.07	* .12	* − .12	** − .14	*** .26	.01
antipathy	.02	.05	*** .18	*** .27	* .14	− .08
agency	.07	.04	.00	.06	*** .26	.10
experiential capacity	.08	.09	* .11	*** .16	*** .29	− .02
unlikability	− .01	.04	*** .19	***.18	.09	.02

TABLE IV

SPEARMAN'S CORRELATION COEFFICIENTS (ρ) FOR THE OUTCOME VARIABLES AND PARTICIPANTS' EXPERIENTIAL BACKGROUND (*victimization* VIA RELATIONAL AGGRESSION) AND ATTITUDINAL DISPOSITIONS (*benevolent* AND *hostile sexism*; *social dominance orientation*; AND *affinity* FOR AND *aversion* TOWARDS ROBOTS). CORRELATIONS WITH ATTITUDES TOWARD ROBOTS ARE COMPUTED USING DATA FROM ONLY PARTICIPANTS SHOWN A ROBOT VICTIM. ASTERISKS DENOTE SIGNIFICANCE (*** DENOTES $p < .001$, ** DENOTES $p < .01$, AND * DENOTES $p < .05$).

may 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 [3], which observed that YouTube commentary on female-gendered robots was more frequently abusive than that regarding male-gendered robots (suggesting 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) and the role they play as abusers, whereas the present work involved random sampling, gave the participants the specific role of bystanders, and had 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 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 observation of abuse may be especially traumatic for those who have previously been subject to relational aggression. *Benevolent sexism* also appears to be predictive 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 connection 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. This suggests these 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, 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, and agency and experiential capacity attributed to, the NAO. Whereas, contrary to what might be expected based on the uncanny valley hypothesis (e.g., [45]), general aversion to robots does not appear to explain people's affect or, rather, disaffection in response to the abuse of a humanoid robot.

A. Limitations

Although our experimental design was well-suited for exploratory assessment of our research questions, its limitations highlight important avenues for further investigation. First, while prior work by Rosenthal-von der Pütten, Tan, and colleagues [9], [10], [5] suggest that the findings here likely extend beyond the Nao to at least Ugobe's animatronic Pleo and Anki's Cozmo platform, future work should explore

whether our results would extend to the broader array of possible robot embodiments (e.g., [46]). Second, future work should investigate whether our results extend beyond physical abuse to other types of socially aggressive behavior (e.g., verbal abuse), abusers of different sizes and genders, and different duration of the video stimuli. (cf. [47]). Third, the present findings are derived from a relatively homogeneous participant sample. Participants were mostly young adults of similar socio-cultural orientation. Future work should investigate whether our results extend to different participant pools (e.g., of different cognitive stages, from different cultural and social contexts). Fourth, future work should use qualitative techniques to better understand how participants felt regarding the abuse, and how they felt the robot should respond. Lastly, future work should compare to a non-agentic control condition (c.p. Rosenthal-von der Pütten's study where an inanimate object (a box) was used as a control[10]). A non-binary agent could also be used to control for victim gendering, although this would face challenges, as people tend to ascribe a binary gender to robots even if provided the option to describe them as androgynous [48]. Despite these opportunities for future work, the present study nevertheless suggests the need to properly respond to robot abuse, and the ways that vulnerable populations can be affected by witnessing abusive interactions.

V. Conclusions

Our findings empirically support the argument that the abuse of a robot can propagate harm to (human) bystanders (cf. [49]), as merely observing 11-seconds of abuse was emotionally distressing to participants. This means that 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). Our findings also highlight that the abuse of robots gendered as female has the potential to serve as a sexist tool for propagating men's social dominance. Specifically, we might anticipate a scenario in which a man abuses a female-presenting robot, with no negative (emotional) consequences to himself, whilst causing harm to witnesses of the interaction.

We propose three key considerations for HRI designers according to our results.

1) Designers should anticipate robot abuse when possible (so that it can be avoided and/or addressed).
2) When abuse can be anticipated, designers should consider whether and how abuse could be avoided. 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 [8]. Robot designers may also be able to adjust the design of robots' physical appearance and social behaviors in order to head off anticipated abuses [50], [51].
3) When prevention is not possible, designers must consider how robots should respond when confronted with abuse, to minimize observer distress, avoid gendered marginalization, and ensure they are not viewed as condoning such actions (cf. [19]). Robots could use empathy to provoke guilt and lower the anger of the offender[52]. However, using this strategy when the robot is perceived as having a marginalized gender identity (e.g. female) risks propogating harmful and/or regressive gender stereotypes [53], [25]Alternatively, robots could actively employ prohibition and shame (such as in the case of public nudity and drinking alcohol) to dissuade abusers from perpetrating public acts of violence [54]; an approach we have pursued in other work grounded in Confucian Role Ethics [55].

These considerations are especially important in light of ongoing diversity issues in computing and robotics, something the UNESCO report explicitly links to gender stereotyping with artificial social agent design/behaviour [53]. Recent work in the field of HRI suggesting that failure to condemn norm-violating actions risks weakening those violated norms [16], further pointing to the need for appropriate responses to abuse, by which we mean those that do not reinforce harmful stereotypes. Moreover, [25] provides initial evidence that responding to abuse can increase robot credibility, and [26] provides cross-cultural replication for these findings. Overall, such confrontation may be critically important in mitigating the adverse impacts to observers and beyond. In order to design appropriate and ethical responses to abuse, qualitative ethnographic research, along with participatory design. Considering that it is specially important to include the perspectives of people of marginalized identities, intersectional feminist research approaches could serve well to accomplish this by approaching the participants with respect towards them and their experiences [56]. The Feminist HRI framework and associated reflexive questions recently laid out by Winkle et al. provide one possible, practical approach for continued HRI work in this area [57].

VI. Acknowledgements

This work was funded in part by Young Investigator award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research.

References

[1] A. De Angeli and S. Brahnam, "I hate you! disinhibition with virtual partners," *Interacting with computers*, vol. 20, no. 3, 2008.
[2] G. Veletsianos, C. Scharber, and A. Doering, "When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent," *Interacting with computers*, 2008.
[3] M. K. Strait, C. Aguillon, V. Contreras, and N. Garcia, "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," in *Int'l Sym. Hum. Rob. Int. Comm. (RO-MAN)*, 2017.
[4] A. C. Curry and V. Rieser, "#metoo alexa: How conversational systems respond to sexual harassment," in *WS on Ethics in NLP*, 2018.
[5] X. Z. Tan, M. Vázquez, E. J. Carter, C. G. Morales, and A. Steinfeld, "Inducing bystander interventions during robot abuse with social mechanisms," in *Int'l Conf. Hum-Rob. Int. (HRI)*, 2018.
[6] D. Küster, A. Swiderska, and D. Gunkel, "I saw it on youtube! how online videos shape perceptions of mind, morality, and fears about robots," *New Media & Society*, vol. 23, no. 11, pp. 3312–3331, 2021.

[7] P. Salvini, G. Ciaravella, W. Yu, G. Ferri, A. Manzi, B. Mazzolai, C. Laschi, S.-R. Oh, and P. Dario, "How safe are service robots in urban environments? bullying a robot," in *Proc. RO-MAN*, 2010.

[8] D. Brščić, H. Kidokoro, Y. Suehiro, and T. Kanda, "Escaping from children's abuse of social robots," in *Proc. HRI*, 2015.

[9] A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler, "An experimental study on emotional reactions towards a robot," *Int'l Journal of Social Robotics*, 2013.

[10] A. M. Rosenthal-Von Der Pütten, F. P. Schulte, S. C. Eimler, S. Sobieraj, L. Hoffmann, S. Maderwald, M. Brand, and N. C. Krämer, "Investigations on empathy towards humans and robots using fmri," *Computers in Human Behavior*, vol. 33, pp. 201–212, 2014.

[11] C. Bartneck and J. Hu, "Exploring the abuse of robots," *Interaction Studies*, vol. 9, no. 3, pp. 415–433, 2008.

[12] R. Sparrow, "Kicking a robot dog," in *Proc. HRI*, 2016.

[13] M. Luria, O. Sheriff, M. Boo, J. Forlizzi, and A. Zoran, "Destruction, catharsis, and emotional release in human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 4, 2020.

[14] L. D. Riek, T.-C. Rabinowitch, B. Chakrabarti, and P. Robinson, "Empathizing with robots: Fellow feeling along the anthropomorphic spectrum," in *Proc. Aff. Computing and Intelligent Interaction*, 2009.

[15] J. Connolly, V. Mocz, N. Salomons, J. Valdez, N. Tsoi, B. Scassellati, and M. Vázquez, "Prompting prosocial human interventions in response to robot mistreatment," in *Proc. HRI*, 2020.

[16] R. B. Jackson and T. Williams, "Language-capable robots may inadvertently weaken human moral norms," in *Proc. HRI*, 2019.

[17] S. Yamada, T. Kanda, and K. Tomita, "An escalating model of children's robot abuse," in *Proc. HRI*, 2020.

[18] S. M. West, M. Whittaker, and K. Crawford, "Discriminating systems: Gender, race and power in ai," *AI Now Institute*, 2019.

[19] A. Schlesinger, K. P. O'Hara, and A. S. Taylor, "Let's talk about race: Identity, chatbots, and ai," in *Proc. CHI*, 2018.

[20] G. Perugia and D. Lisy, "Robot's gendering trouble: A scoping review of gendering humanoid robots and its effects on hri," *arXiv preprint arXiv:2207.01130*, 2022.

[21] T. Nomura, "Robots and gender," *Gender and the Genome*, 2017.

[22] D. Cameron and E. C. Collins, "Children's reasoning on robots and gender," in *Int'l Conference on Social Robotics*, 2020.

[23] R. B. Jackson, T. Williams, and N. Smith, "Exploring the role of gender in perceptions of robotic noncompliance," in *Proc. HRI*, 2020.

[24] M. Strait, P. Briggs, and M. Scheutz, "Gender, more so than age, modulates positive perceptions of language-based human-robot interactions," in *Int'l Sym. new frontiers in human robot interaction*, 2015.

[25] K. Winkle, G. I. Melsión, D. McMillan, and I. Leite, "Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots," in *2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[26] K. Winkle, R. B. Jackson, G. I. Melsión, D. Brščić, I. Leite, and T. Williams, "Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study," in *Proc. HRI*, 2022.

[27] S. A. Basow, K. F. Cahill, J. E. Phelan, K. Longshore, and A. McGillicuddy-DeLisi, "Perceptions of relational and physical aggression among college students: Effects of gender of perpetrator, target, and perceiver," *Psychology of Women Quarterly*, 2007.

[28] S. Stewart-Williams, "Gender, the Perception of Aggression, and the Overestimation of Gender Bias,"

[29] M. B. Harris and K. Knight-Bohnhoff, "Gender and aggression I: Perceptions of aggression," *Sex Roles*, vol. 35, pp. 1–25, July 1996.

[30] C. Williams, D. S. Richardson, G. S. Hammock, and A. S. Janit, "Perceptions of physical and psychological aggression in close relationships: A review," *Aggression and Violent Behavior*, vol. 17, pp. 489–494, Nov. 2012.

[31] J. M. Grant, L. A. Motter, J. Tanis, *et al.*, "Injustice at every turn: A report of the national transgender discrimination survey," 2011.

[32] S. A. Fahlman, K. B. Mercer-Lynn, D. B. Flora, and J. D. Eastwood, "Development and validation of the multidimensional state boredom scale," *Assessment*, 2013.

[33] J. P. Abel, C. L. Buff, and S. A. Burr, "Social media and the fear of missing out: Scale development and assessment," *Journal of Business & Economics Research (JBER)*, vol. 14, no. 1, pp. 33–44, 2016.

[34] L. E. Watkins, R. C. Maldonado, and D. DiLillo, "The cyber aggression in relationships scale: A new multidimensional measure of technology-based intimate partner aggression," *Assessment*, 2018.

[35] P. Glick and S. T. Fiske, "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism.," *J. Pers.& Soc. Psych.*, 1996.

[36] F. Pratto, J. Sidanius, L. M. Stallworth, and B. F. Malle, "Social dominance orientation: A personality variable predicting social and political attitudes.," *Jour. personality and social psychology*, 1994.

[37] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, "Measurement of negative attitudes toward robots," *Interaction Studies*, 2006.

[38] N. Ezer, A. D. Fisk, and W. A. Rogers, "Attitudinal and intentional acceptance of domestic robots by younger and older adults," in *Proc. universal access in human-computer interaction*, 2009.

[39] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales.," *Journal of personality and social psychology*, 1988.

[40] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *science*, vol. 315, no. 5812, pp. 619–619, 2007.

[41] M. M. De Graaf and B. F. Malle, "People's explanations of robot behavior subtly reveal mental state inferences," in *Proc. HRI*, 2019.

[42] P. H. Kahn Jr, T. Kanda, H. Ishiguro, N. G. Freier, R. L. Severson, B. T. Gill, J. H. Ruckert, and S. Shen, ""robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot.," *Developmental psychology*, 2012.

[43] M. Kwon, M. F. Jung, and R. A. Knepper, "Human expectations of social robots," in *Proc. HRI*, 2016.

[44] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of social issues*, 2000.

[45] M. Strait, L. Vujovic, V. Floerke, M. Scheutz, and H. Urry, "Too much humanness for human-robot interaction: exposure to highly humanlike robots elicits aversive responding in observers," in *Proc. ACM conference on human factors in computing systems*, 2015.

[46] E. Phillips, X. Zhao, D. Ullman, and B. F. Malle, "What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database," in *Proc. HRI*, 2018.

[47] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using robots to moderate team conflict: the case of repairing violations," in *Proc. HRI*, 2015.

[48] G. Perugia, S. Guidi, M. Bicchi, and O. Parlangeli, "The shape of our bias: Perceived age and gender in the humanoid robots of the abot database," in *Proc. HRI*, 2022.

[49] B. Whitby, "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents," *Interacting with Computers*, 2008.

[50] H. Ku, J. J. Choi, S. Lee, S. Jang, and W. Do, "Designing shelly, a robot capable of assessing and restraining children's robot abusing behaviors," in *Companion Proc. HRI*, 2018.

[51] M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe, and R. Tow, "Experiences with sparky, a social robot," in *Socially intelligent agents*, 2002.

[52] H. Chin, L. W. Molefi, and M. Y. Yi, "Empathy is all you need: How a conversational agent should respond to verbal abuse," in *Proc. CHI Conference on Human Factors in Computing Systems*, 2020.

[53] M. West, R. Kraut, and H. Ei Chew, "I'd blush if I could: closing gender divides in digital skills through education," tech. rep., 2019.

[54] K. Mamak, "Should violence against robots be banned?," *International Journal of Social Robotics*, vol. 14, 06 2022.

[55] Q. Zhu, T. Williams, B. Jackson, and R. Wen, "Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective," *Science and Engineering Ethics*, vol. 26, pp. 2511–2526, 2020.

[56] K. Bhopal, "Gender, identity and experience: Researching marginalised groups," *Women's Studies Int'l Forum*, 2010.

[57] K. Winkle, D. McMillan, M. Arnelid, K. Harrison, M. Balaam, E. Johnson, and I. Leite, "Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 72–82, 2023.