Robots Need the Ability to Navigate Abusive Interactions

Hideki Garcia Goo*, Katie Winkle†, Tom Williams‡, and Megan Strait§
*University of Twente, Enschede, Netherlands. Email: h.garciagoo@utwente.nl
†KTH Royal Institute of Technology, Stockholm, Sweden. Email: winkle@kth.se
‡Colorado School of Mines, Golden, United States. Email: twilliams@mines.edu
§University of Texas Rio Grande Valley, Edinburg, United States. Email: megan.strait@utrgv.edu

***Abstract*—Researchers are seeing more and more cases of abusive disinhibition towards robots in public realms. Because robots embody gendered identities, poor navigation of antisocial dynamics may reinforce or exacerbate gender-based violence. It is essential that robots deployed in social settings be able to recognize and respond to abuse in a way that minimises ethical risk. Enabling this capability requires designers to first understand the risk posed by abuse of robots, and hence how humans perceive robot-directed abuse. To that end, we experimentally investigated 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.**

***Index Terms*—HRi, robot abuse**

I. Introduction

The increasingly 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 verbal and physical abuse toward artificial social agents accumulating across academic and public domains [1]–[6].

Much of existing research on robot abuse has focused on the potential for robot abuse to impact those perpetrating that abuse (typically negatively [7], [8], although cf. [9]). However, the impacts of abuse, and a victim's response to it, extend not only to abusers, but to bystanders as well [10]–[14]. Moreover, the effects of abusing a robot – as well as witnessing a robot's abuse – likely extend beyond a single interaction (e.g., [8]) and may risk normalization [15] – or escalation [16] – of that behavior. Consequently, **agents unable to navigate antisocial dynamics risk replicating, reinforcing, and exacerbating extant social inequities** [17].

Since people ascribe robots a gender, not only on the basis of stereotypic cues in a robot's presentation [18]–[20]), voices and names [21]–[23]), but also due to robot-unique factors like physical morphology [24], even to robots not intentionally gendered [25]. In turn, this enables robots to similarly evoke and reinforce gendered stereotypes in a complex way that interacts with interactants' gender identities [25]–[27]. Thus, it is critical for robot designers to have a nuanced understanding of these complex gender-mediated perceptions and their implications.

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, and building upon the seminal research by Rosenthal-von der Pütten and colleagues [11], [12], 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. We then showed participants videos of these depictions and assessed 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.

II. Method

Based on work by Astrid Rosenthal-von der Pütten and colleagues [11], [12], we showed participants 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.

Because of prior associations between participants' gender and their evaluations of human-robot interactions (e.g., [25]), and given that the perception of abuse itself is gendered (e.g., [28]), we quasi-manipulated participants' **gender** via binary categorization of their self-identification as a *man* or

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

with a *marginalized* identity (e.g., genderfluid, nonbinary, woman). All study materials and procedures were approved by the Institutional Review Board at the University of Texas Rio Grande Valley under protocol IRB-20-0379.

A. Participants

In total, 482 participants consented, after excluding those that failed the attention check ($n = 27$) or quit before completing their session ($n = 33$), data from 417 remained. Of these 417 participants, 63% identified as women, 35% identified as men, and 2% identified with nonbinary identities. 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).

B. Design

Stimuli: 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: the perpetrator (1) thrusts the victim against the table; (2) suffocates the victim with a plastic bag; and (3) suffocates the victim with a rope around their neck. In all videos, both the perpetrator and victim are positioned facing away from the camera to avoid differences in facial affect and 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 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 [29], [30]). 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

Experiential background and attitudinal dispositions: Using the (Cyber) Aggression in Relationships Scale [31], we assessed the frequency at which participants experienced *victimization* (via psychological aggression) in the past year, and, using the Ambivalent Sexism Inventory [32] and SDO-16 [33], we assessed participants' *benevolent sexism*, *hostile sexism*, and *social dominance orientation*. In addition, we derived two constructs reflecting one's *affinity* for and *aversion* to robotic technologies, using the Negative Attitudes towards Robots Scale [34] and the Robot Acceptance Scale [35].

Effects of the manipulations: Using the Positive and Negative Affect Schedule [36] and the Mind Perception Scale [37], 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 ([11], [12]), 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* (Item was reverse-scored.) and was *unaffected*[0] 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

The orientations of the constructs' global means (i.e., average across all samples; see Tab. 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. 106 participants watched the video

portraying the victim as a female robot, 102 watched a male robot, 102 watched a female human, and 106 watched a male human.

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. 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. Tab. 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.

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* ($ps < .001$), see Tab. II.

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 ($p < .001$), see Tab. 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, see Tab. 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.

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 Tab. 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. *Hostile sexism* and *social dominance orientation*, on the other hand, appear predictive of insensitivity to the abuse; both were associated with participants' antipathy 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 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.

IV. Discussion

To understand the risks posed by abusive human-robot interactions, the present work explored the socio-emotional impacts of witnessing a robot's abuse.

A. Implications

Witnessing the abuse of robots is distressing. Consistent with the observations by Rosenthal-von der Pütten [11], [12], and by Tan [14], 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. Though, also consistent with [11], [12], 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).

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. 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.

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. Moreover, **men are particularly antipathetic to a robot's abuse** (or at least they portray themselves to be).

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 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.

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. Also, 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 (e.g., [38]), 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.

B. Design Considerations

The findings outlined above, mean 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 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.

This clearly motivates three key considerations for HRI designers. First, designers must attempt to anticipate robot abuse when possible, 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. [39]). 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 [6]. Additional approaches include assuming abuse of a robot will occur and adjusting its physical design and social behavior to provide negative feedback [40], [41], and strategically employing shame and guilt to dissuade abusers from perpetrating further acts of violence [42].

These considerations are especially important in light of the UNESCO report [43], 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 [15]. Moreover, [44] provides initial evidence that responding to abuse can increase robot credibility. Overall, such confrontation may be critically important in mitigating the adverse impacts to observers and beyond.

C. Limitations

Reproduction of this research with different robotic platforms, forms of abuse, and participant pools (e.g., of different cognitive stages, from different cultural and social contexts) is recommended to test its generalizability. The results might also change if the participants witness this interaction in a lab instead of online.

V. Conclusions

The present findings suggest, in particular, 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

[1] A. De Angeli and S. Brahnam, "I hate you! disinhibition with virtual partners," *Interacting with computers*, vol. 20, no. 3, pp. 302–310, 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*, vol. 20, no. 3, pp. 292–301, 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 *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 1418–1423.

[4] A. C. Curry and V. Rieser, "# metoo alexa: How conversational systems respond to sexual harassment," in *Proceedings of the second acl workshop on ethics in natural language processing*, 2018, pp. 7–14.

[5] 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 *19th International Symposium in Robot and Human Interactive Communication*. IEEE, 2010, pp. 1–7.

[6] D. Brščić, H. Kidokoro, Y. Suehiro, and T. Kanda, "Escaping from children's abuse of social robots," in *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*, 2015, pp. 59–66.

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

[8] R. Sparrow, "Kicking a robot dog," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 229–229.

[9] 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, pp. 1–19, 2020.

[10] L. D. Riek, T.-C. Rabinowitch, B. Chakrabarti, and P. Robinson, "Empathizing with robots: Fellow feeling along the anthropomorphic spectrum," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–6.

[11] 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," *International Journal of Social Robotics*, vol. 5, no. 1, pp. 17–34, 2013.

[12] 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.

[13] 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 *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 211–220.

[14] X. Z. Tan, M. Vázquez, E. J. Carter, C. G. Morales, and A. Steinfeld, "Inducing bystander interventions during robot abuse with social mechanisms," in *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 2018, pp. 169–177.

[15] R. B. Jackson and T. Williams, "Language-capable robots may inadvertently weaken human moral norms," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 401–410.

[16] S. Yamada, T. Kanda, and K. Tomita, "An escalating model of children's robot abuse," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 191–199.

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

[18] F. Eyssel and F. Hegel, "(s) he's got the look: Gender stereotyping of robots," *Journal of Applied Social Psychology*, vol. 42, no. 9, pp. 2213–2230, 2012.

[19] N. Fitter, M. Strait, E. Bisbee, M. Matarić, and L. Takayama, "You're wigging me out! is personalization of telepresence robots strictly positive?" in *2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2021.

[20] D. Cameron and E. C. Collins, "Children's reasoning on robots and gender," in *12th International Conference on Social Robotics*. Springer, 2020.

[21] D. Kuchenbrandt, M. Häring, J. Eichberg, F. Eyssel, and E. André, "Keep an eye on the task! how gender typicality of tasks influence human–robot interactions," *International Journal of Social Robotics*, vol. 6, no. 3, pp. 417–427, 2014.

[22] C. McGinn and I. Torre, "Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 211–221.

[23] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.

[24] J. Bernotat, F. Eyssel, and J. Sachse, "Shape it–the influence of robot body shape on gender perception in robots," in *International Conference on Social Robotics*. Springer, 2017, pp. 75–84.

[25] T. Nomura, "Robots and gender," *Gender and the Genome*, vol. 1, no. 1, pp. 18–25, 2017.

[26] R. B. Jackson, T. Williams, and N. Smith, "Exploring the role of gender in perceptions of robotic noncompliance," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 559–567.

[27] M. Strait, P. Briggs, and M. Scheutz, "Gender, more so than age, modulates positive perceptions of language-based human-robot interactions," in *4th international symposium on new frontiers in human robot interaction*, 2015, pp. 21–22.

[28] 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*, vol. 31, no. 1, pp. 85–95, 2007.

[29] S. A. Fahlman, K. B. Mercer-Lynn, D. B. Flora, and J. D. Eastwood, "Development and validation of the multidimensional state boredom scale," *Assessment*, vol. 20, no. 1, pp. 68–85, 2013.

[30] 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.

[31] 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*, vol. 25, no. 5, pp. 608–626, 2018.

[32] P. Glick and S. T. Fiske, "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism." *Journal of personality and social psychology*, vol. 70, no. 3, p. 491, 1996.

[33] F. Pratto, J. Sidanius, L. M. Stallworth, and B. F. Malle, "Social dominance orientation: A personality variable predicting social and political attitudes." *Journal of personality and social psychology*, vol. 67, no. 4, p. 741, 1994.

[34] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, "Measurement of negative attitudes toward robots," *Interaction Studies*, vol. 7, no. 3, pp. 437–454, 2006.

[35] N. Ezer, A. D. Fisk, and W. A. Rogers, "Attitudinal and intentional acceptance of domestic robots by younger and older adults," in *International conference on universal access in human-computer interaction*. Springer, 2009, pp. 39–48.

[36] 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*, vol. 54, no. 6, p. 1063, 1988.

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

[38] 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 *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 3593–3602.

[39] A. Schlesinger, K. P. O'Hara, and A. S. Taylor, "Let's talk about race: Identity, chatbots, and ai," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–14.

[40] 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 of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 161–162.

[41] M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe, and R. Tow, "Experiences with sparky, a social robot," in *Socially intelligent agents*. Springer, 2002, pp. 173–180.

[42] K. Darling, "Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects," in *Robot law*. Edward Elgar Publishing, 2016.

[43] 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. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[44] 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.

APPENDIX

	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 * $\Rightarrow p < .05$).

	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
experiential 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 * $\Rightarrow p < .05$).

	participant gender				
	men	*marg.*	$M_d \pm SE$	t	d
negative 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	$.01 \pm .29$	$.02 \pm .32$	$-.01 \pm .03$	$-.43$	$-.04$
hostile sexism	$-.08 \pm .40$	$-.27 \pm .35$	$.18 \pm .04$	*** 4.77	.49
soc. dom. orient.	$-.44 \pm .34$	$-.52 \pm .31$	$.07 \pm .03$	* 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.

	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 $p < .05$).