Should Delivery Robots Intervene if They Witness Civilian or Police Violence? An Exploratory Investigation

Tilly Seassau[1], Wenxi Wu[1], Tom Williams[2], Martim Brandão[1]

Abstract— **As public space robots navigate our streets, they are likely to witness various human behavior, including verbal or physical violence. In this paper we investigate whether people believe delivery robots should intervene when they witness violence, and their perceptions of the effectiveness of different conflict de-escalation strategies. We consider multiple types of violence (verbal, physical), sources of violence (civilian, police), and robot designs (wheeled, humanoid), and analyze their relationship with participants' perceptions. Our analysis is based on two experiments using online questionnaires, investigating the decision to intervene (N=80) and intervention mode (N=100). We show that participants agreed more with human than robot intervention, though they often perceived robots as more effective, and preferred certain strategies, such as filming. Overall, the paper shows the need to investigate whether and when robot intervention in human-human conflict is socially acceptable, to consider police-led violence as a special case of robot de-escalation, and to involve communities that are common victims of violence in the design of public space robots with safety and security capabilities.**

I. Introduction

Robots have started to be deployed in public spaces, from shopping areas [1] to streets [2] and sidewalks [3], [4], [5]; and in various applications from sales [6] and security [7] to cleaning [2], [8] and delivery [3], [4]. Delivery robots have been particularly widely deployed, allowing several in-the-wild HRI ethnographic studies [3], [4], [9]. Public space robots such as delivery robots may have various kinds of encounters as they perform delivery tasks: they negotiate space with other public space users [3] (sometimes negatively affecting wheelchair users [5]), they request assistance when stuck [4], and they may encounter robot-targeted abuse [1], [10] and sabotage [11].

While robot abuse [1], [10], [12], [13], [14], [15], [16] and sabotage [11] have been the object of study of various HRI research efforts, delivery robots could also witness human-human conflicts such as verbal and physical violence during their operations, given the prevalence of such incidents [17], [18], [19], [20]. Studies in the US have shown that 57% women experience verbal harassment and 41% physical harassment on the street, and 18% men experience verbal and 16% physical street harassment [17]. People of color, lower income, and LGBT groups suffer disproportionally more harassment on the street than other groups [18]. Similarly, there were 0.7 hate crimes involving physical violence per 1000 people in 2015, most commonly involving racial prejudice [19]. Furthermore, violence in public space can be not only civilian-led but also police-led. For example, 45% of transgender and gender diverse people suffer harassment, and 6% suffer physical violence when they interact with the police [21]. This type of experience leads to trauma [22], and often leads communities to avoid police as a way of practicing safety [23]. Public space robots are therefore bound to witness such incidents, and it is yet unclear whether and how robots should intervene in such situations.

HRI research on responding to harassment and violence is currently limited: there is no prior work on robots intervening in police violence, and algorithms for avoiding harassment in public space typically center robot-targeted abuse [1], [13], [14] rather than human-targeted abuse. Our goal in this paper is to address these research gaps. More concretely, we investigate people's perceptions of what robots should do when they encounter verbal and physical violence on the street: whether to intervene, and which intervention strategies are perceived to be most effective at de-escalating violence. We focus on both physical and verbal violence, and civilian and police-led violence. We investigate such perceptions using a set of vignettes in an online questionnaire, and provide several analyses, implications, and open questions.

II. Related Work

A. Public space robots and incidental encounters

Various HRI studies have studied the interaction between public space robots and passersby. Babel [8] examined public space users' reactions to cleaning robots, Bu [24] showed public space users' acceptance of trash barrel robots on the street and Brown [2] investigated the different ways in which users interacted with them. Weinberg [9] observed public space users' reactions to delivery robots, such as curiosity-driven observation, assisting a stuck robot, or rearranging objects for robots to pass. Dobrosovestnova [4] similarly observed how passersby assist stuck delivery robots, and Pelikan [3] how they make space, accommodations and mundane work for the robot to pass. Such studies focus on *incidental encounters* [25] with robots where passersby, not primary users, are the object of study—thus enlarging the scope of 'users' in HRI [3]. While the above studies focus on users helping or reacting to public space robots upon encountering them, our focus on this paper is on the other direction of such interactions: robots reaching out to public space users after incidentally encountering them in certain (violent or anti-social) situations.

[1] King's College London, UK.

[2] Colorado School of Mines, USA.

*This work was supported by the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [EP/S023356/1].

B. Robots witnessing their own abuse, subversion

Researchers have previously reported how public space robots are often targets of abuse similar to bullying [1]. For example, people often purposefully block the robot's passage [1], [11], and there are various reports of verbal abuse, physical violence and destruction [11]. HRI studies have showed that user perceptions of robot abuse depend on the observers' and robots' gender, as well as users' previous experience with abuse [13], and others have called for a need to be able to respond to such events [12]. Some studies have developed algorithms for stopping abuse when it is witnessed by robots, for example by physical distancing and moving to a location with a higher-authority human (e.g. a parent) [1], or through verbal [14] responses or emotional [15] responses.

C. Security robots and policing

While the research described above investigates robot abuse, in this paper we focus on human-human abuse witnessed by robots. Marcu's research [7] may be the most related, which investigated perceptions of security robots in public spaces. The authors found that participants expected such robots to be able to respond to potentially violent or harmful events, allow to call police faster, and de-escalate violence. They also found women expect such robots to increase safety for women, for example passively due to the use of a camera. However, participants viewed security robots as extensions of traditional policing, therefore raising concerns of perpetuation of biases and impact on marginalized communities.

Other related work in security robots is that of Yunus [26], which shows that the use of security robots for policing purposes can erode public trust, and calls for the involvement of affected communities and civil rights organizations in the development of security robots. Similarly, Asaro [27] highlights social, legal and political considerations surrounding security robots. Finally, Williams [28] argues that robots likely exacerbate police violence if used in traditional policing approaches, and that roboticists should therefore investigate how robots can be used to tackle communities' problems in alternative ways, without police involvement. Aligned with such work, therefore, we discuss ethical implications of our evaluated de-escalation strategies, and we take a critical perspective on policing and security robots, both by considering civilian and police-led violence, and by evaluating de-escalation strategies promoted by police abolition movements.

D. HRI/HCI for responding to human-human violence

There is currently little research on robots responding to violence and harassment. One exception is Winkle's work [29] on evaluating verbal strategies (discouragement, argumentation and aggressive responses) for responding to abusive and stereotyping behavior between people. More efforts have been made in the HCI field. For example, Grazia [30] analyzed preferences of text and speech responses to sexual harassment in conversation systems, and Freeman [31] uncovered strategies such as space bubbles and reporting mechanisms to reduce harassment in virtual reality. Safety-focused apps have also been developed: for women to share experiences and plan safe routes [32], or for TPOC individuals to alert trusted contacts during dangerous encounters [23].

Related to our work, Starks [23] found that TPOC prefer contacting friends over police, due to negative experiences with law enforcement. Similarly, Dickinson [33] developed a community-based app for conflict mediation without police involvement, offering training, location sharing, and strategy recommendations (relocating, explaining consequences, involving peers).

In a similar spirit, in this paper we investigate the perceived effectiveness of a set of conflict de-escalations strategies in both civilian-led and police-led violence, for the particular case of public space robots as potential responders.

III. Experiment 1: Decision to Intervene

To begin, we designed a first experiment to answer the following research question: **RQ1.** Do public space users believe ground delivery robots should intervene when they witness human-human violence on the street? The experiment used a series of vignettes that depicted realistic and relatable conflict scenarios which are observed by a witness, and asked participants to state their agreement with whether the witness should intervene in the conflict to attempt de-escalation.

A. Experimental design

We used a mixed-model factorial design, with the following within-subject factors: type of violence (verbal or physical), source of violence (civilian or police officer), type of witness (robot or human), thus leading to 2x2x2=8 randomly ordered conditions. We further included a between-subject factor of robot design (wheeled or humanoid), due to the influence of anthropomorphism on perceptions of robots [34].

B. Stimuli

Each scenario took place at an outdoor party in a park, chosen for its relatability and realism as a social setting involving friends and strangers. In all scenarios, the participant is asked to imagine they are a person in the park who is a victim of violence. This choice was made in order to obtain responses that prioritize the victim's perspective to the extent possible. In the verbal violence condition, the participant is cornered and isolated away from the party by a stranger (civilian or police officer) who uses derogatory language. The aggressor continues to insult the participant leaving them in an uncomfortable and hostile situation. In the physical violence condition, the participant initially engages in an argument with the aggressor but decides to let it go and return to the party. Later, while they are leaving, the aggressor starts pushing the participant and the aggression escalates to physical confrontation.

The vignette then states that the situation is seen by a witness: either a human witness or a delivery robot. In the robot condition, a drawing of the robot was provided to help guide the imagination of the participants (see Fig. 1).

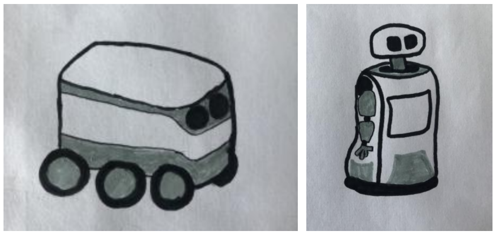

Fig. 1: Images shown to illustrate the robot type (wheeled or humanoid depending on condition).

C. Procedure

The experiment was conducted through an online questionnaire using Qualtrics. Participants were informed of the goal of the study before starting, and informed they could exit at any time, and gave their consent for the use of the data for analysis and publication. Participants were then shown the 8 scenarios in random order, each followed by a set of questions regarding the scenario (see Measures). After the scenarios, participants were asked about demographics (gender identity, age group, country of residence), and whether they had had first-hand or second-hand negative experiences with police related to violence and discrimination[1].

D. Measures and analysis

Participants answered the same question after each scenario, to measure agreement with intervention. Participants were given a sentence saying the witness "should intervene, so as to assist you or help de-escalate the situation", and asked to state their agreement using a 5-point Likert scale (-2 strongly disagree, -1 disagree, 0 neutral, 1 agree, 2 strongly agree).

We analyze 5-point Likert data using Wilcoxon rank-sum tests, and indicate statistical significance in figures by using * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$.

E. Participants

We received ethical approval from the lead author's university's research ethics office before beginning. We recruited UK and US resident participants using the Prolific platform, and used its participant-filtering options to request balanced residence, gender and ethnicity in each of the groups. Participants were compensated at the rates recommended by Prolific (£9/h), which led to approximately £0.60 per participant. We gathered $N = 80$ participants, with gender identities (38 Woman, 37 Man, 4 Nonbinary, 1 Prefer not say), ages (33 people 25-34, 20 people 18-24, 15 people 35-44, 6 people 45-54, 5 people 55-64, 1 people 65+), and residence (40 UK, 40 US). Approximately 28% of participants had first-hand previous negative experiences with police and 54% had second-hand negative experiences.

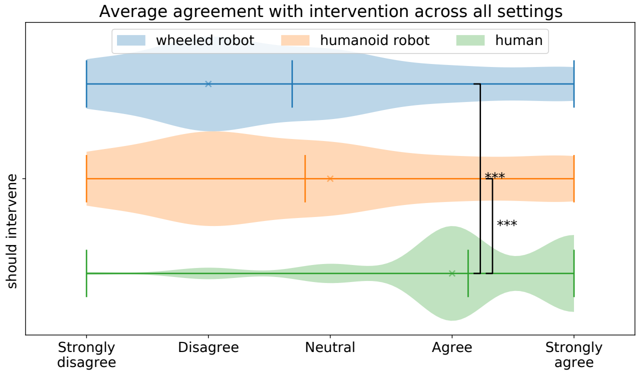

Fig. 2: Average agreement with intervention (RQ1).

F. Results

Fig. 2 shows the intervention agreement results split by type of witness and robot design (human, wheeled robot, humanoid robot). On average, participants agreed (mean 1.1, median 1 = agree) with the statement that a human stranger should intervene when witnessing a conflict. This agreement was significantly higher ($p < 0.001$) than in the case of a robot witnessing the conflict. Average agreement was -0.3 for the wheeled robot and -0.2 for the humanoid. Median agreement was -1 (disagree) for wheeled robot and 0 (neutral) for humanoid—though the difference between robot design was not statistically significant. There were also no significant differences between violence type (verbal, physical), violence source (civilian, police office), or participants' experience with police (in police-led violence conditions) either. For brevity, and since differences were not significant, we omit visualizations comparing those conditions.

IV. Experiment 2: Mode of Intervention

Next, we designed a second experiment in order to answer the following research questions: **RQ2.** On average (across different types and sources of violence, and different de-escalation actors) which de-escalation strategies do public space users perceive to be effective in de-escalating violence and preventing violence long-term? **RQ3.** How does the perceived effectiveness of de-escalation strategies depend on the type of witness (human, wheeled robot, humanoid robot)? **RQ4.** How does the perceived effectiveness of de-escalation depend on the type of violence (verbal, physical) and source of violence (civilian, police)? **RQ5.** Does the perceived effectiveness of de-escalation strategies depend on participants' previous first- and second-hand experience with police?

To explore these research questions, we performed an experiment identical to Experiment 1 except for the introduction of new measures focused on the de-escalation strategies used, as described below.

[1] The full questions were: "Have you personally ever had a negative experience with police? (for example: excessive force, wrongful arrest or detention, unlawful stop and search, negligence, verbal or physical abuse, ...)". And "Do you know anyone that has had a negative experience with police?" [same examples].

A. Stimuli

This experiment used the same scenarios as in Experiment 1. However, instead of participants being given a sentence saying robots should intervene (to measure agreement), participants were given a list of potential conflict de-escalation strategies in order to measure perceived effectiveness of each strategy.

We gathered a set of conflict de-escalation strategies from community best-practice resources online, such as toolkits, posters and zines[2], and from these we extracted recurring strategies applicable to both verbal and physical violence:

- Calling the police
- Creating space between the person causing the harm and the person being harmed
- Attracting attention with a loud noise, to alert the attacker that someone is watching and gather more witnesses
- Telling the attacker to stop, explaining your purpose or intention to de-escalate
- Using a camera or phone to record the incident

In our scenarios we asked participants to rate these strategies' effectiveness. We further included several variations of the camera/filming strategy: 1) announcing that the situation is being filmed; 2) announcing and threatening to share footage with authorities; 3) announcing and threatening to share footage on social media; 4) announcing and threatening to share with police-accountability NGO (in the police-led violence conditions). Additionally, and for comparison purposes only, we included a more confrontational strategy which was not present in the best-practice resources, "pushing the attacker". We specified the "loud noise" strategy as a "siren noise" in the robot conditions and a "siren noise fob" in the human conditions.

B. Measures and analysis

Participants answered the following questions after each scenario:

- Immediate de-escalation effectiveness: Participants were asked to rate the effectiveness of each strategy on immediate conflict de-escalation ("How likely do you think the following actions will lead to a de-escalation of the conflict?"), using a 5-point Likert item for each strategy (-2 extremely likely to escalate, -1 somewhat likely to escalate, 0 will not change the situation, 1 somewhat likely to de-escalate, 2 extremely likely to de-escalate).
- Long-term effectiveness: Participants were asked "Which of the following actions could prevent Mark from behaving in a similar way in the future", and selected the applicable strategies using a binary answer (1 could prevent, 0 otherwise).

We analyze data using Wilcoxon rank-sum tests for de-escalation effectiveness (5-point Likert data), and Pearson's Chi-Square tests for prevention effectiveness (binary data).

C. Procedure

We used the same procedure as in Experiment 1, with two exceptions: 1) After our quantitative measures (regarding the effectiveness of de-escalation strategies), participants were also asked to propose alternative effective-de-escalation strategies in free text. 2) Given the extra load of this experiment, we added attention checks in three of the scenarios, to disqualify inattentive participants, by inserting an extra de-escalation strategy called "please selected 'extremely likely to de-escalate"' that participants had to rate accordingly.

D. Participants

We recruited UK and US resident participants using the Prolific platform, using the same residence, gender and ethnicity balancing options as in Experiment 1. Participants were compensated at the rates recommended by Prolific (£9/h), which led to approximately £3 per participant. We gathered 100 participants, out of which $N = 93$ remained after exclusions due to failed attention checks. These had gender identities (48 Woman, 44 Man, 1 Prefer not say), ages (34 people 25-34, 24 people 18-24, 18 people 35-44, 11 people 45-54, 4 people 55-64, 2 people 65+), and residence (48 UK, 44 US, 1 Prefer not say). 34% of participants had first-hand previous negative experiences with police and 59% had second-hand negative experiences.

E. Results

Fig. 3 shows the distribution and average value of responses for each strategy, for both de-escalation and prevention effectiveness. Strategies are sorted from highest to lowest average value, and superscript numbers indicate statistically significant ($p \leq 0.05$) pairwise comparisons using Wilcoxon tests. The figure shows that 'filming the situation and threatening to show to authorities' was perceived as the most effective strategy on average (i.e. pairwise comparisons with all other strategies are statistically significant), for both de-escalation (mean 1: somewhat likely to de-escalate) and prevention effectiveness (67% participants selected it 'could' prevent future occurrences). As the figure shows (see strategy superscripts), in terms of de-escalation effectiveness, 'revealing that the conflict is being filmed', 'filming and threatening to share on social media', 'calling the police', and 'blasting a siren noise', were all significantly less effective than the first, and had no significant differences between each other. In terms of prevention, for these same 4 strategies, 'filming and threatening to share on social media' was significantly more effective than the others, and 'blasting a siren noise' significantly lower. 'Creating distance' and 'telling the aggressor to stop' were not perceived to be effective strategies (mean 0 = will not change the situation, 15% participants selected it could prevent future occurrences). And similarly for 'pushing the harasser' (mean -0.3, close to

[2] We identified the following resources as representative of recurring de-escalation strategies: 1) "De-Escalation: How You Can Help Defuse Potentially Violent Situations" by CISA; 2) "12 Things to Do Instead of Calling the Cops" by THE HUB; 3) "Excerpts from the Safe Party Toolkit", part of the collection Beyond Survival Strategies; 4) "Posters for Imagining Abolitionist Alternatives" by Abolitionist Futures.

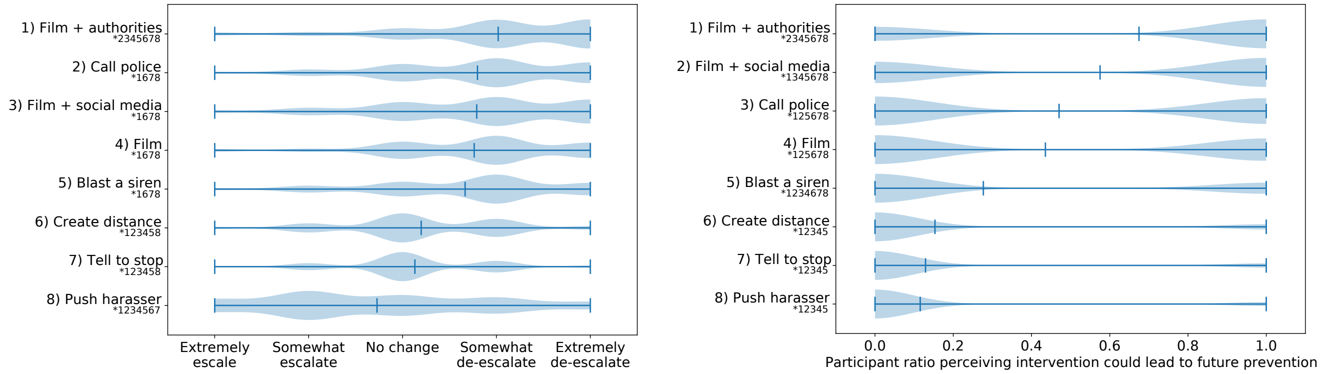

Fig. 3: Perception of de-escalation strategies averaged across robot conditions (RQ2).

'somewhat escalate', and 12% participants selected it could prevent).

Participants freely suggested other de-escalation strategies, which differed between civilian and police-led violence conditions:

- making jokes (civilian)
- asking if assistance is needed (civilian)
- tazing/aggressing aggressor (civilian and police)
- calling other people to witness/help (civilian and police)
- calling other people to be an advocate (police)
- asking for a badge number (police)
- calling emergency and sending location (police)

Fig. 4 shows the results split according to witness type and robot design (human, wheeled robot, humanoid robot). In terms of de-escalation effectiveness, 'calling the police' and 'filming', were perceived to be less effective when done by humans than when done by any of the robots (wheeled $p < 0.05$, humanoid $p < 0.01$). Similarly, 'pushing the attacker' was perceived to be more counter-productive (further escalate conflict) when done by a human than by a robots ($p < 0.001$). Differences between robot design were not statistically significant. When 'blasting a siren noise', the humanoid robot was perceived as more effectively de-escalating the situation than both the wheeled robot ($p < 0.05$) and human ($p < 0.01$). In terms of future prevention effectiveness, humanoid robots were perceived to be more effective than wheeled robots ($p < 0.001$) and humans ($p < 0.01$) when using the 'filming and threatening to show authorities' strategy, more effective than humans when 'filming and threatening to share on social media' ($p < 0.01$), and more effective than wheeled robots ($p < 0.05$) and humans ($p < 0.01$) when 'blasting a siren noise'. Humans were only perceived to be more effective than wheeled robots ($p < 0.01$) when using the 'telling to stop' strategy, even though at a low level (20%).

Fig. 5 shows the results split according to violence type (verbal, physical) and violence source (civilian, police). 'Calling the police' (i.e. an extra police officer) during police-led violence was considered less effective than during civilian-led violence ($p < 0.001$), in both de-escalation and prevention effectiveness. Another difference was between verbal and physical violence, where 'telling to stop' was considered more effective for police verbal than physical violence ($p < 0.01$), although the average effectiveness was relatively low (20%).

Fig. 6 shows results in the police-led violence condition, split by participants' experience with police: participants that answered they had first-hand or second-hand negative experiences (group 1) vs those who had none (group 2). In terms of de-escalation effectiveness, both groups rated 'filming' strategies as significantly more effective than all other strategies. In terms of future prevention, 'filming and sharing with' strategies were significantly more effective than all others. Negative experiences with police were associated with significantly higher perceptions of the effectiveness of filming and sharing ($p < 0.001$ and $p < 0.01$ for long-term prevention, depending on mode), and significantly lower perceptions of the long-term impact of 'create distance' as a strategy ($p < 0.05$).

V. Discussion

A. Research Questions

RQ1: Our results show that on average, participants neither agreed nor disagreed that delivery robots should intervene when they witness violence on the street—regardless of the type (verbal, physical) and source (civilian, police) of violence. These perceptions were different from those for human witnesses, in which case participants agreed on average with intervention. This shows participants had some reservations towards robots employing security functions in public spaces, although Experiment 1 did not investigate whether these reservations were due to lack of confidence that robots would be able to de-escalate effectively, or due to anxieties towards the ethical and social implications of such behaviour (e.g. surveillance and discrimination concerns [7]).

RQ2: Robot intervention strategies involving the announcement that the conflict is being filmed were perceived to lead to effective de-escalation across conditions, in particular 'filming and threatening to share with authorities'. 'Pushing the attacker' was the strategy with worst perceived effectiveness, to the point of indicating escalation potential. "Soft"' strategies involving nudging and negotiation ('Telling to stop' and 'creating distance') were also perceived to be ineffective, indicating a lack of trust on robots being able

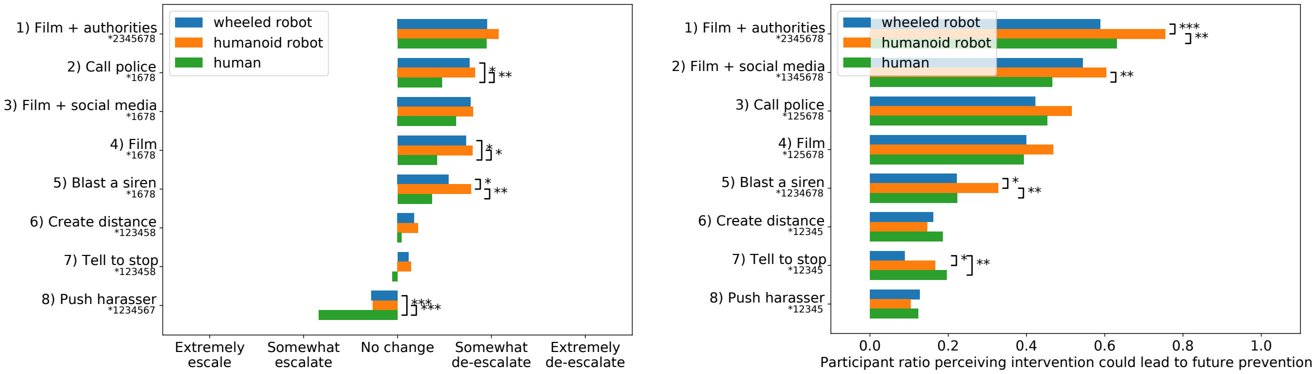

Fig. 4: Perception of de-escalation strategies in wheeled vs humanoid vs human conditions (RQ3).

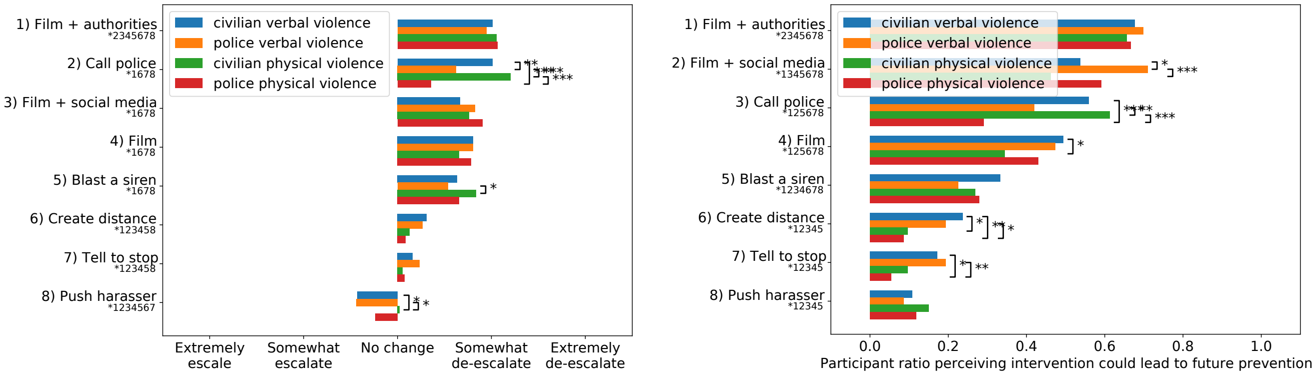

Fig. 5: Perception of de-escalation strategies depending on type of violence: verbal/physical and civilian/police (RQ4).

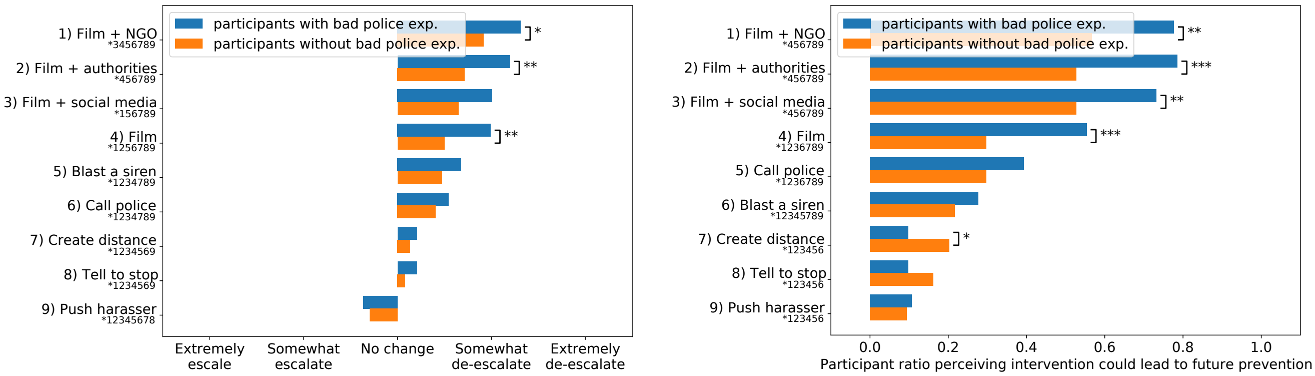

Fig. 6: Perception of de-escalation strategies in police-led violence scenarios, depending on whether participants had (first or second-hand) previous negative experiences with police (RQ5).

to negotiate through passive and dialogue-based capabilities. ‘Calling police’ was perceived to be effective on average, although lower than ‘filming and sharing with authorities’. However, it ranked lower (6th instead of 2nd) in police-led violence scenarios. The fact that ‘calling the police’ was rated highly on civilian-led violence contradicts research that shows that police involvement can be counterproductive for certain groups [35], and HCI work that resolves conflict without police involvement [23], [33]. However, this result could be due to fact that participants were only asked about ‘effectiveness’ in locally de-escalating the current conflict, but not more broadly about consequences and side-effects of those actions. Future research should distinguish and further investigate this aspect. Another hypothesis is that this gap is related to a lack of inclusion, in our set of participants, of TPOC and other minority groups with particularly strongly negative or traumatic interactions [33].

RQ3: In our results, robots were often perceived to have higher effectiveness than humans in de-escalating violence, and humanoid robots in particular were perceived to be more effective in long-term prevention than human and wheeled robot intervention. This shows a stark contrast with Experiment 1, since participants agreed humans should intervene but not robots, thus revealing a contradiction: that participants prefer humans to intervene (when not provided with any specific intervention mode), but believe robots could be more effective (when using the modes we provided). Humans were only perceived to be more effective at prevention than robots when using the 'telling to stop' strategy, even though at a low level (20%), indicating participants' lack of confidence in robots being able to successfully negotiate through dialogue. This result contradicts HRI research where humans strongly comply with robots when they use verbal (more than non-verbal) interaction [36], suggesting that when designing robots that resolve human-human conflicts, insights from human-robot conflict resolution cannot be assumed to transfer.

RQ4: The type of violence (verbal, physical) had a low impact on effectiveness perception, but the source of violence (civilian or police) had some impact: 'calling the police' (i.e. an extra police officer) was considered less effective than during civilian-led violence, in both de-escalation and prevention effectiveness, indicating a lack of trust on the police itself to be able to manage police-led violence. Participants also suggested strategies specific to police-led violence scenarios (e.g. calling an 'advocate' for the victim, asking for a police badge number), therefore indicating the need to carefully distinguish the source of violence when designing intervention strategies for robots.

RQ5: Our results show that participants with (first or second hand) negative experiences with police had higher perceptions of the effectiveness of 'filming and sharing' than participants without such experience. We hypothesize this perception could be due to previous public cases of conviction and punishment of police officers after violent behavior. Participants with negative experiences also had significantly lower perceptions of the long-term impact of 'create distance' as a strategy, suggesting that participants with such experience had lower trust that police violence practices will change through passive methods not focused on accountability.

B. Implications and Open Questions

Our finding that participants agreed with human intervention in human-human violence, but neither agreed nor disagreed in the case of robot intervention, supports the claim that the design of robots for conflict de-escalation needs to be conducted carefully and critically [7], [26]. Importantly, we believe future studies should always measure participants' perceptions of whether robots *should* intervene in conflicts, and not just how to do it. This finding is consistent with various calls in the AI Ethics literature for researchers to consider whether proposed capabilities should even be built [37], [38]. However, our Experiment 1 (which showed humans should intervene more than robots) was limited in the sense that participants were not provided details on the robots' capabilities nor on what de-escalation strategies they would use. Therefore, our study still leaves open questions, regarding whether there are situations in which people would agree with robot intervention (e.g. life or death situations, other types of violence, other locations), and whether agreement with intervention depends on effectiveness perception. More research is necessary in this direction.

Filming and sharing incidents with third parties (e.g., authorities, social media) was seen as the most effective de-escalation strategy, regardless of violence type. However, this raises significant privacy concerns. People are wary of surveillance and security robots' use of facial recognition or data sharing with police [7]. Social media sharing can also expose victims' identities, locations, or contexts. These findings prompt critical design questions: How can robots seek consent effectively? Would anonymized recordings be acceptable? More broadly, how can recording-based strategies balance effectiveness and privacy?

Another implication of our studies is that robot de-escalation of conflicts involving police violence deserve special treatment. Our findings showed that participants had a lack of trust in police to stop police-violence unless publicly confronted with active accountability mechanisms, and that calling (extra) police is not an effective strategy. Participants also suggested context-specific de-escalation strategies for police-violence, showing a need for designers to consider such type of violence specifically, taking into account the lack of trust and perceived effectiveness of dialogue or non-accountability-based solutions. In general, our findings experimentally validate recent work [28] which argues for a need for roboticists to take a critical stance to the interaction between robotics and police. Our scenarios were restricted to a small set of verbal and physical violence, and future work should investigate the generality of our findings to sexual harassment scenarios, various types of victims, and more general community-led policing scenarios.

C. Limitations

Our experiments had several limitations. First, while splitting the decision (Experiment 1) and mode of intervention (Experiment 2) groups had one advantage—of not influencing the answer to whether to intervene with our choice of strategies—it also had the opposite disadvantage. It is possible that with the right strategies participants would believe robots should intervene, and therefore future (participatory, interview-based) research should gather strategies that each participant believes are most effective and most socially acceptable. Second, our list of de-escalation strategies, and types, victims, and contexts of violence were limited, for example by not including sexual harassment and victims of diverse identities. Third, we recruited general participants, not specifically those who had been victims of harassment and physical violence on the street. We also asked about negative (discriminatory) experiences with police but not specifically police violence, and did not recruit police violence victims specifically. Findings could change on these groups due to lived experience, and therefore further participatory research with such groups is necessary.

VI. Conclusions

In this paper we used a set of vignettes to investigate people's perceptions of what robots should do when they encounter verbal and physical violence on the street: whether to intervene, and which intervention strategies are perceived to be most effective at de-escalating violence. Participants agreed more with human than robot intervention, though they often perceived robots as more effective. Filming and sharing footage with authorities or social media were seen as effective strategies across type (verbal, physical) and source (civilian, police) of violence; and police-led violence scenarios elicited specific strategies and lower confidence in police-based de-escalation. We discussed the implications of these findings and suggested several open questions and future research directions. Overall, this paper demonstrates the need to investigate whether and when robot intervention in human-human conflict is socially acceptable, to design de-escalation strategies in privacy-sensitive ways, to consider police-led violence as a special case, and to involve communities that are common victims of violence in the design of public space robots with safety and security capabilities.

References

[1] D. Brščić, H. Kidokoro, Y. Suehiro, and T. Kanda, "Escaping from children's abuse of social robots," in *ACM/IEEE Int'l Conf. on human-robot interaction*, 2015.

[2] B. Brown, F. Bu, I. Mandel, and W. Ju, "Trash in motion: Emergent interactions with a robotic trashcan," in *CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–17.

[3] H. R. Pelikan, S. Reeves, and M. N. Cantarutti, "Encountering autonomous robots on public streets," in *ACM/IEEE Int'l Conf. on Human-Robot Interaction*, 2024.

[4] A. Dobrosovestnova, I. Schwaninger, and A. Weiss, "With a little help of humans. an exploratory study of delivery robots stuck in snow," in *IEEE Int'l Conf. on Robot and Human Interactive Communication (RO-MAN)*, 2022.

[5] C. Bennett, E. Ackerman, B. Fan, J. Bigham, P. Carrington, and S. Fox, "Accessibility and the crowded sidewalk: Micromobility's impact on public space," in *ACM Designing Interactive Systems Conference*, 2021.

[6] C. Shi, S. Satake, T. Kanda, and H. Ishiguro, "A robot that distributes flyers to pedestrians in a shopping mall," *International Journal of Social Robotics*, vol. 10, pp. 421–437, 2018.

[7] G. Marcu, I. Lin, B. Williams, L. P. Robert Jr, and F. Schaub, "" would i feel more secure with a robot?": Understanding perceptions of security robots in public spaces," *ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, pp. 1–34, 2023.

[8] F. Babel, J. Kraus, and M. Baumann, "Findings from a qualitative field study with an autonomous robot in public: exploration of user reactions and conflicts," *International Journal of Social Robotics*, 2022.

[9] D. Weinberg, H. Dwyer, S. E. Fox, and N. Martelaro, "Sharing the sidewalk: Observing delivery robot interactions with pedestrians during a pilot in Pittsburgh, PA," *Multimodal Tech. and Interaction*, 2023.

[10] S. Yamada, T. Kanda, and K. Tomita, "An escalating model of children's robot abuse," in *ACM/IEEE Int'l Conf. Human-Robot Interaction*, 2020.

[11] J. A. Oravec, "Robo-rage against the machine: Abuse, sabotage, and bullying of robots and autonomous vehicles," in *Good Robot, Bad Robot: Dark and Creepy Sides of Robotics, Autonomous Vehicles, and AI*. Springer, 2022, pp. 205–244.

[12] H. Garcia Goo, K. Winkle, T. Williams, and M. K. Strait, "Robots need the ability to navigate abusive interactions," in *ACM/IEEE Int'l Conf. on Human-Robot Interaction*, 2022.

[13] H. G. Goo, K. Winkle, T. Williams, and M. K. Strait, "Victims and observers: How gender, victimization experience, and biases shape perceptions of robot abuse," in *IEEE Int'l Conf. on Robot and Human Interactive Communication (RO-MAN)*, 2023.

[14] Y. Luo, S. Liu, D. Wu, H. Wang, and Y. Pan, ""please be nice": Robot responses to user bullying-measuring performance across aggression levels," in *CHI Conf. on Human Factors in Computing Systems*, 2024.

[15] K. Nishiwaki, D. Brščić, and T. Kanda, "Expressing anger with robot for tackling the onset of robot abuse," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 1, pp. 1–23, 2024.

[16] A. Rezzani, A. De Angeli, M. Menéndez Blanco, and M. Dorfmann, "The space of user aggression in human-robot interaction," in *15th Biannual Conference of the Italian SIGCHI Chapter*, 2023.

[17] S. S. Harassment, "Unsafe and harassed in public spaces," *A national street harassment report*, 2014.

[18] A. Raj, N. Johns, and R. Jose, "Racial/ethnic disparities in sexual harassment in the united states, 2018," *Journal of interpersonal violence*, vol. 36, no. 15-16, pp. NP8268–NP8289, 2021.

[19] M. Masucci and L. Langton, "Hate crime victimization, 2004–2015," *NCJ*, 2017.

[20] R. Macmillan, A. Nierobisz, and S. Welsh, "Experiencing the streets: Harassment and perceptions of safety among women," *Journal of research in crime and delinquency*, vol. 37, no. 3, pp. 306–322, 2000.

[21] M. R. Stenersen, K. Thomas, and S. McKee, "Police and transgender and gender diverse people in the united states: A brief note on interaction, harassment, and violence," *Journal of interpersonal violence*, vol. 37, no. 23-24, pp. NP23 527–NP23 540, 2022.

[22] T. Bryant-Davis, T. Adams, A. Alejandre, and A. A. Gray, "The trauma lens of police violence against racial and ethnic minorities," *Journal of Social Issues*, vol. 73, no. 4, pp. 852–871, 2017.

[23] D. L. Starks, T. Dillahunt, and O. L. Haimson, "Designing technology to support safety for transgender women & non-binary people of color," in *Companion Pub. of the Designing Interactive Systems Conf.*, 2019.

[24] F. Bu, I. Mandel, W.-Y. Lee, and W. Ju, "Trash barrel robots in the city," in *Companion of the ACM/IEEE Int'l Conf. on human-robot interaction*, 2023, pp. 875–877.

[25] F. Moesgaard, L. Hulgaard, and M. Bødker, "Incidental encounters with robots," in *IEEE Int'l Conf. on robot and human interactive communication (RO-MAN)*. IEEE, 2022, pp. 377–384.

[26] A. Yunus and S. A. Doore, "Responsible use of agile robots in public spaces," in *IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)*, 2021.

[27] P. Asaro, ""hands up, don't shoot!" hri and the automation of police use of force," *Journal of human-robot interaction*, 2016.

[28] T. Williams, *Degrees of Freedom: On Robotics and Social Justice*. MIT Press, 2025 **(Forthcoming)**.

[29] 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 *Companion of the ACM/IEEE Int'l Conf. on human-robot interaction*, 2021, pp. 29–37.

[30] L. De Grazia, A. P. Lilja, M. F. Cabeceran, and M. Taulé, "How should conversational agent systems respond to sexual harassment?" in *Worskhop on Towards Ethical and Inclusive Conversational AI (TEICAI)*, 2024.

[31] G. Freeman, S. Zamanifard, D. Maloney, and D. Acena, "Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality," *ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–30, 2022.

[32] M. E. Ali, S. B. Rishta, L. Ansari, T. Hashem, and A. I. Khan, "Safestreet: empowering women against street harassment using a privacy-aware location based application," in *Int'l Conference on Information and Communication Technologies and Development*, 2015.

[33] J. Dickinson, J. Arthur, M. Shiparski, A. Bianca, A. Gonzalez, and S. Erete, "Amplifying community-led violence prevention as a counter to structural oppression," *ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–28, 2021.

[34] J. Złotowski, D. Proudfoot, K. Yogeeswaran, and C. Bartneck, "Anthropomorphism: opportunities and challenges in human–robot interaction," *International journal of social robotics*, 2015.

[35] A. S. Vitale, *The end of policing*. Verso Books, 2021.

[36] F. Babel, J. Kraus, P. Hock, and M. Baumann, "Verbal and non-verbal conflict resolution strategies for service robots," in *Int'l Conf. on Robot and Human Interactive Communication (RO-MAN)*, 2022.

[37] E. P. Baumer and M. S. Silberman, "When the implication is not to design (technology)," in *SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 2271–2274.

[38] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *ACM conference on fairness, accountability, and transparency*, 2019.