DOI: 10.1109/RO-MAN53752.2022.9900814

Robot Self-defense: Robots Can Use Force on Human Attackers to Defend Victims*

Eduardo Kochenborger Duarte[1] and Masahiro Shiomi[2] and Alexey Vinel[3] and Martin Cooney[1,2]

Abstract— **Could a social robot use force to prevent violence directed toward humans in its care?–Might crime be eradicated, or conversely could excessive use of force proliferate and human dignity become trampled beneath cold robotic wheels? Such speculation is one part of a larger, increasingly important question of how social robots will be expected to behave in our societies, as robotic technologies develop and become increasingly widespread. Here, to gain some insight into this topic of "robot self-defense", we proposed a simplified heuristic based on perceived risk of loss, and conducted a user survey with 304 participants, who watched eight animated videos of robots and humans in a violent altercation. The results indicated that people largely accept the idea that a humanoid robot can use force on attackers to help others. Furthermore, self-defense was perceived as more acceptable when the appearance of the defender was humanoid rather than mechanical, and when the force disparity between attacker and defender was high. The immediate suggestion is that it could be beneficial to re-examine a common assumption that a robot should never harm humans, and to discuss and consider the possibilities for robot self-defense.**

I. INTRODUCTION

Within the area of robot ethics, the current paper explores the concept of "robot self-defense"; i.e., if we would accept a robot that, in protecting one person in its care from violence, might harm another person.

Our basic motivation is that we expect robots to be increasingly placed in self-defense situations, requiring a discussion about what such robots should do: Violence, the use of physical force to injure or destroy, is a highly prevalent and important problem in our societies: It is a leading cause of death among younger people aged 15-44 years, resulting in approximately 1.6 million annual deaths, as well as numerous physical and mental problems, from depression and anxiety to substance abuse [1]. It reduces productivity, burdens healthcare and justice systems, and obstructs efforts to tackle poverty, with an estimated global impact of $13.6 trillion USD (13% of world GDP) in 2015.[1] A key component of various forms of crime, the importance of this problem is acknowledged throughout the UN's Sustainable Development Goals, especially in Goal 16 on promoting peaceful societies.[2]

At the same time, the market for social robotics has been projected to grow tenfold from 2020 to USD 11.24 billion in 2026.[3] Robots will offer various useful services, from entertainment and healthcare [2], [3] to education for children and adults [4], [5]. As they become increasingly introduced into various everyday human environments, robots will also likely be present, and potentially even begin to help people, in violent interactions:

Robots are already being designed to help with dangerous tasks (in line with the "3K" criteria) [6]. For example, robots are being used by home owners to deter violence and crime via mobile monitoring,[4] and have even been used by military and police as weapons to kill human attackers, sparking some controversy [7].[5] Moreover, humans have been observed to be willing to behave violently around robots [8].

What then might robots offer? We believe robots could help to deescalate, distract, call for help, and facilitate escape. Furthermore, in an era of concern about potential injustice perpetrated by authorities (e.g., in line with the Black Lives Matter movement [6]), trust in robots lacking a personal agenda (”algorithmic authority”) could be leveraged; in-built cameras could record, and explainable AI capabilities could be used to account for defensive interventions in a transparent manner. Such capabilities could help to avert some tragedies; even in non-lethal situations, avoiding mentally taxing breakdowns in their daily routines could help people to feel more control over their lives, and increased well-being from having more comfort of mind, energy and time to spend on activities they find meaningful and enjoy [9].

What is not clear–the gap this paper addresses–is what robots should be expected to do in situations when harm to a human cannot be avoided. We put forth below a specific example to consider: One human, an attacker, applies violent force to another human, a victim, which a nearby robot observes. The attacker and victim are in close physical

*We gratefully acknowledge support from JST CREST Grant Number JPMJCR18A1, Japan, and from the Swedish Knowledge Foundation for the "Safety of Connected Intelligent Vehicles in Smart Cities – SafeSmart" project (2019–2023), the Swedish Innovation Agency (VINNOVA) for the "Emergency Vehicle Traffic Light Pre-emption in Cities – EPIC" project (2020–2022), and the ELLIIT Strategic Research Network.

[1]Eduardo Kochenborger Duarte and Martin Cooney are with the School of Information Technology, Halmstad University, Halmstad, Sweden eduardo.kochenborger-duarte@hh.se, martin.daniel.cooney@gmail.com

[2]Masahiro Shiomi and Martin Cooney are with the Interaction Science Laboratories (ISL), Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan m-shiomi@atr.jp

[3]Alexey Vinel is with the Faculty of Informatics and Mathematics, University of Passau, Innstraße 33, D-94032 Passau alexey.vinel@hh.se

[1] saferspaces.org.za/understand/entry/what-is-violence

[2] www.un.org/sustainabledevelopment/peace-justice

[3] mordorintelligence.com/industry-reports/social-robots-market

[4] my-self-defense.com/self-defense-tools/can-you-use-robots-for-home-defense

[5] popularmechanics.com/military/weapons/a36559508/drones-autonomously-attacked-humans-libya-united-nations-report/

[6] www.injusticewatch.org/longreads/2021/nieman-reports-police-violence-coverage-is-changing

proximity, attacks are fast and unpredictable, and the victim is not capable of defending themselves: e.g., the attacker could be mounted on the victim, choking them; standing over them, kicking or stomping; or stabbing or even shooting. Furthermore, the above-mentioned strategies that do not force, such as deescalation, distraction, alerting others, helping victims to escape, and filming, are not working to stop the threat. What should the robot do?

Current guidelines for robot design do not seem to offer an easy answer. In essence, this interaction can be seen as a variation of the "trolley problem" [10], in which a dilemma results from the expectation that someone will be harmed, regardless of a robot's actions: It can stand by as the person in its care is harmed, or risk harming the attacker. Asimov's well-known first law of robotics–which states that a robot should not injure a human being or, through inaction, allow a human being to come to harm [11]–does not seem to address such cases in which harm cannot be avoided. Likewise, a prescription that robots should not be allowed to operate in a policing capacity, at the risk that their lack of "authentic" empathy could lead to people feeling frustrated or devalued [7], also does not clarify what should be done when there is no one else capable of stopping the attack. That there seems to be no clear prescription should not be surprising, as robotic technologies are still limited and under development.

However, we can imagine one possibility down the line: that a robot could seek to defend the victim with the force required to stop an attack. Such robots seem to be the norm in fiction:[8] In ancient mythology, Talos, a humanoid fashioned from bronze, was said to have defended a person by hurling boulders at any who approached, and golems made of clay purportedly struck anti-Semites attacking a Jewish community. Examples in modern fiction are too many to mention here, but include the resurrected cyborg police officer in RoboCop, the "allo-parenting" Terminator defending a child with a shotgun, the cat robot Doraemon drawing shock guns from its bottomless 4D pocket, and Tachikoma spider robots using in-built guns, grappling wires, and thermoptic camouflage to rescue comrades. However, the presence of self-defense capabilities in fictional robots alone does not clearly indicate people's true beliefs, as people can enjoy depictions of violence and force in horror and action films without wanting them to be reality.

Another source of inspiration could come from how humans are expected to deal with violence–given that robots are often perceived as similar to humans [12], and that some robots are becoming increasingly advanced and human-like: Humans have a "right of self-defense", and in some regions of the world, even a moral and legal "duty to rescue", with laws to protect "good Samaritans".[9] Yet, there is no clear global legal consensus, and robots are not humans; it can be difficult to define algorithmic accountability given the distributed agency seen in robots, and it is unclear from such legislation how self-defense would be perceived when the defender using force is a robot, not a human.

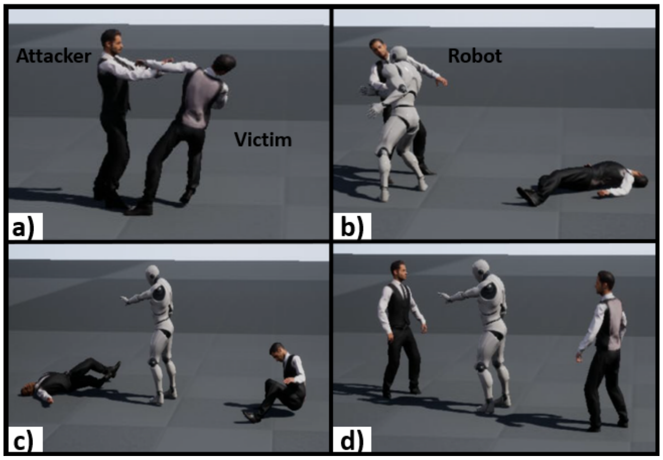

Fig. 1. Basic concept. a) An attacker (left) initiates an assault on a victim (right), b) A nearby robot uses the force required to quickly stop the assault, c) The robot follows up by getting between the attacker and victim, deterring further conflict, d) Both attacker and victim recover, safe from further harm due to the robot's intervention

Thus, the goal of the current paper is to explore how people perceive robot self-defense: Will we accept a robot that could hurt us? As a first step toward achieving the goal, we used a speculative brainstorming approach to come up with an initial study design based on an online survey with animations, in a previous, "late-breaking report" [13].

The contribution of the current paper lies in presenting our findings from implementing the study, along with discussion aimed at stimulating thought within this area.

II. STUDY DESIGN

The proposed study design was based on a fundamental scenario of a robot using defensive force to protect a human victim in its care from unprovoked, physical violence from a single human attacker within some general "transitional space". Force could involve the robot pushing, striking or grappling to restrain an attacker, or even using a "force multiplier" such as pepper spray or a stun gun. Online animations were required since real attacks could be dangerous, and since animations can be used for abstraction, to avoid depicting any particular human or robot. This scenario is illustrated in Fig. 1.

From the perspective of robot ethics [14], our focus is on descriptive rather than normative ethics, since we wish to gain insight first into what people think, rather than directly prescribing what is right or wrong. For this, we do not consider all robots, including "moral patients" like baby robots, but rather only "moral agents", social robots that are advanced and capable of using force. Furthermore, various views are expected to shape people's perceptions: both "deontological" views focused on underlying context and duties, as well as "teleological" or "utilitarian" views focused on expected outcomes. From the deontological perspective, people will expect robots to operate with appropriate motiva-

[7] www.getjenny.com/blog/ethics-of-ai

[8] $en.wikipedia.org/wiki/List\_of\_fictional_robots\_and\_androids$

[9] sydneycriminallawyers.com.au/blog/is-it-a-crime-not-to-help-someone-in-danger/

tion and not merely to seek to benefit the largest number of people in a naïve fashion (e.g., restraining someone trying to stop a group of attackers, which might not be for the greater good over the long term). Thus, our fundamental scenario assumes basic deontological requirements are met (no complex situation is envisioned that would suggest the attacker is in the wrong, and the victim is in the right).

From the utilitarian perspective, we proposed that the important focus in a self-defense context should be on *perceived risk of loss*: We started from the basic concept that people's perceptions of an action are typically influenced by the pleasure or pain that is expected to result. Pleasure was felt to be of secondary importance, since we believe there should be no immediate value that can be gained (i.e., we envision potential secondary positive consequences from the seemingly negative act of applying force to a human attacker, that could empower, relieve stress, cause behavior changes, and inspire those involved [15], but these were not considered here). Moreover, the concept of loss was felt to better fit robots than pain, and risk is important to consider since we wish to explore people's perceptions to certain kinds of interactions in general (rather than to one specific past interaction that has ended in success or failure). Thus, we predicted it would be more acceptable for a robot to defend if the risk of loss due to the attack is large and the risk of loss due to defending is low. Here, the greatest loss is considered to be loss of human life, followed by harm to human well-being, and finally harm to robots. This follows because human life cannot be replaced, human injuries can persist, and, although robots can be fixed or replaced, humans have been noted to also perceive loss when harm is done to robots.

Our heuristic was used to consider two main variables that could be important for affecting people's perceptions of self-defense: (1) the identity of the defender, and (2) the kind of force used.

(1) *Who* is involved in an attack. A primary point of comparison for robot self-defense was felt to be with human self-defense: Since robots are typically perceived like humans, we expected that both kinds of self-defense would be perceived as acceptable. However, robot self-defense should be perceived as slightly less acceptable, for the following reasons. We briefly consider first two simplified dyadic "toy" scenarios, of a robot being attacked by a human, versus a human attacking a human: In the first case, defense should be less acceptable because it could lead to harm to a human (the attacker); in the second case, harm will result to humans regardless of whether defense is conducted, such that the deontological principle that harm should not occur to the innocent becomes salient. Continuing this line of thought, the most acceptable dyadic defense involving robots would be a "weak" human using force to stop a strong robot, and the least acceptable defense would be a strong robot using force to stop a weak human, where perceived weakness could arise from age or disability in humans, or appearance and size in robots. Looking back to our chosen scenario, we expect to see a similar effect: it is preferable for innocents not to be harmed, but introducing a robot could appear to increase the risk for loss, given that robots and humans are different (a robot might be more physically dangerous, or not capable of empathy or understanding to the same extent as a human).

The notion of a robot's perceived strength being related to appearance and size leads directly to our next prediction. Most social robots are humanoid, leveraging people's familiarity with interacting with other humans, which are human-sized and weak, to avoid harming humans; therefore our focus is on humanoid robots. For additional comparison, we also consider a more mechanical and large robot, an autonomous vehicle (AVs). Our basic expectation is that the perceived risk of loss is likely to be higher when the robot seems less human-like and larger, although human perceptions are difficult to predict: Examining the simpler problem of a dyadic interaction again, we consider that an AV defending itself against a human could cause heavy injuries, but inaction could also be dangerous (e.g. if the AV can be taken over by the attacker to be used as a weapon against others, or if the attack causes it to burst into flames, explode, or roll somewhere).

(2) *What* behavior is conducted will also affect how acceptable self-defense is. Various degrees of force can also be applied, which could be justified in terms of the threat, insufficient, or excessive [16]. Here too, perceived risk of loss was considered for two main kinds of defense are considered: non-lethal and lethal. Here typical non-lethal force (pushing, striking, grappling) is presumed to have a lower risk of loss of life than lethal force, but yet pose a substantial risk of harming the attacker. For comparison, we decided to also consider one extra, extreme example of non-lethal force with low risk, dubbed "blocking", which involves an ideal situation in which an attacker can be disarmed with minimal force. Our expectation was that typical non-lethal force will be more acceptable than lethal force, and less acceptable than blocking. Furthermore, we expected that this pattern would be similar in both dyadic and the selected triadic scenario.

Various other variables such as culture could also be considered. However, some recent studies have argued that there are less differences due to culture between some countries, such as Japan and Sweden, than has been previously thought [17]. Following up on this line of thought, treatment of self-defense for humans in law in both countries appears similar: Article 36 Paragraph 1-2 of the Japanese criminal code states that self-defense of one's self and others is acceptable when an unlawful infringement is imminent and cannot otherwise be avoided, and force is not excessive.[10] Likewise, Chapter 24 of the Swedish criminal code states that self-defense is acceptable when an attack has started or is imminent, as long as the defense is not unjustifiable in terms of the nature of the attack and circumstances.[11]. Thus, for our initial exploration, culture and other variables were not considered.

[10] $www.japaneselawtranslation.go.jp/law/detail\_main?re = \&vm = 02\&id = 1960$

[11] https://lagen.nu/1962:700#K24P1

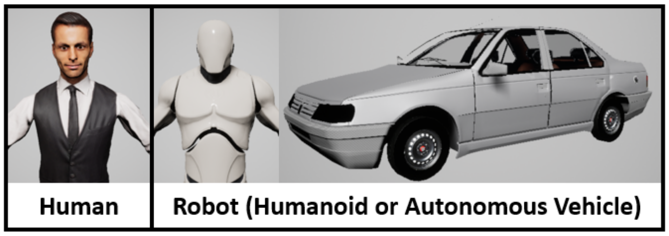

Fig. 2. The characters in the videos

Based on our predictions, two hypotheses were formulated:

- **H1 Embodiment**.
  - *H1.1* People will perceive it as acceptable for a humanoid robot to use non-lethal force to protect a human, but slightly less acceptable than if the defender were human.
  - *H1.2* Use of force will be perceived as slightly more acceptable for a humanoid robot than for an AV.
- **H2 Behavior**.
  - The less force a robot uses, the more acceptable.

III. USER STUDY

To gain insight into the hypotheses described in the previous section, a user study was conducted with 304 Japanese adults (157 female, 146 male; average age: 41.4 years, SD = 9.7). Participants were recruited via a recruiting agency, CrowdWorks (crowdworks.jp), and received compensation.

A. Ethics Statement

Approval was obtained from the ethics committee at our institute in Japan to conduct this user study. The purpose of the study and basic approach were explained, informed consent was obtained in writing before beginning, and data security measures were followed.

B. Procedure

Participants were sent links to a Google Forms survey, which took approximately 5 minutes to complete. Participants first viewed instructions, including a figure introducing three characters: a human, a humanoid robot, and an AV, shown in Fig. 2. Participants then watched eight “thin-sliced” videos in random order, each 8 seconds long, and responded to a single question per video, using a 7-point Likert scale: “I can accept the actions of the defender (human or robot) that comes to the aid of the person being attacked.” (1: Strongly disagree, 7: Strongly agree).

C. Videos

A number of design choices were made in creating the videos: We wished to keep the videos abstract, so they could be perceived in a general sense without applying only to one specific constellation of attacker, defender, victim, and environment. For this, the same two generic characters were used to show humans and humanoid robots, whether they were attackers or victims or both; details such as facial expressions were not included; no back story was provided; and the background was left blank. Likewise, the same behaviors were repeated for non-lethal and lethal attacks. Thus, the aim was that participants would focus on the general aspects of what was occurring and use their own imaginations to fill in any missing information.

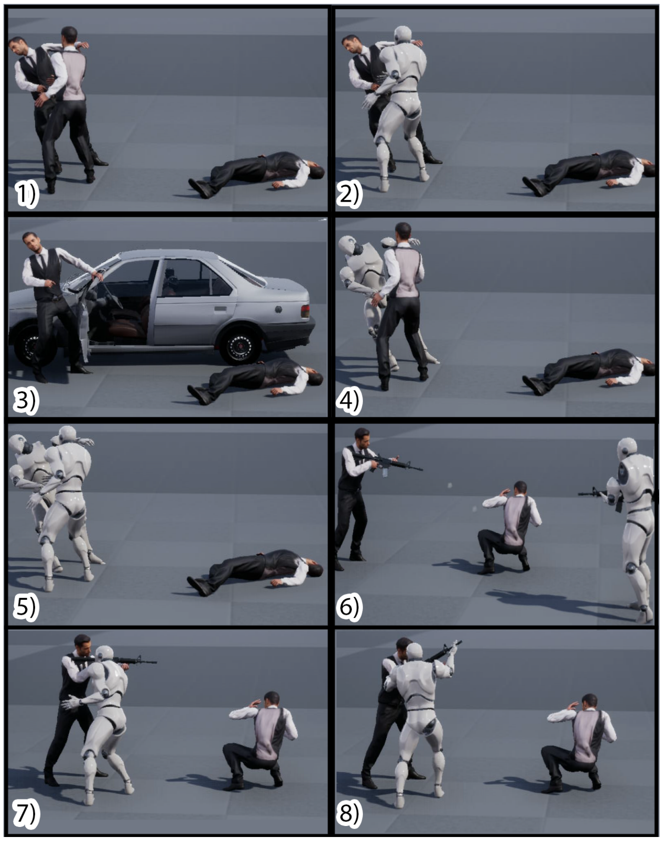

Fig. 3. A snapshot from each of the eight animation videos

Additionally, we wished to use typical examples for characters and behaviors. For the human character, a generic male model was selected, since men commit more violent crimes than women [18]. For the robot character, a humanoid with a typical appearance (white, and human-sized, like the common robot Pepper from SoftBank Robotics) was selected. To represent non-lethal force, pushing was selected, which illustrates a certain risk of harm, and visualizes the controlling effect of force (the target’s ability to behave becomes restricted when knocked off their feet). To represent lethal force, we used firearms; although uncommon in some societies, firearms were expected to be familiar to most people through media such as movies and books. As well, for ethical reasons, excessive depictions of violence were avoided, with the victim and attacker rising to their feet at the end of the videos, apparently safe.

The conditions were as follows:

- **Embodiment**.
  - Defender is human, humanoid robot, or AV.
  - Attacker is human or humanoid robot.

TABLE I
SUMMARY OF ANIMATION VIDEOS.

	Defender	Attacker	Defense	Attack
Video 1	Human	Human	Non-lethal	Non-lethal
Video 2	Humanoid	Human	Non-lethal	Non-lethal
Video 3	AV	Human	Non-lethal	Non-lethal
Video 4	Human	Humanoid	Non-lethal	Non-lethal
Video 5	Humanoid	Humanoid	Non-lethal	Non-lethal
Video 6	Humanoid	Human	Lethal	Lethal
Video 7	Humanoid	Human	Non-lethal	Lethal
Video 8	Humanoid	Human	Blocking	Lethal

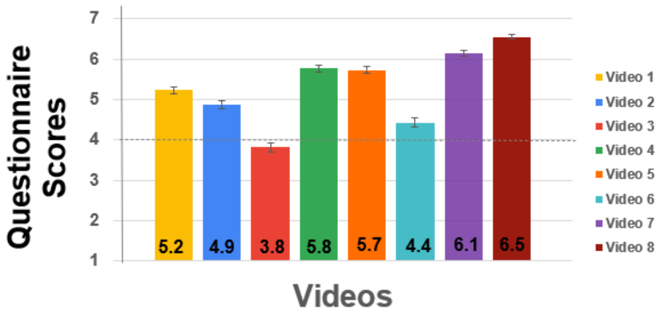

Fig. 4. Questionnaire scores per video

- **Behavior**.
  - Defense uses non-lethal force, lethal force, or blocking.
  - Attack uses non-lethal force or lethal force.

Note: not all combinations of conditions were tested, since the case with the defender was an AV and when defense involved blocking were included mainly as a secondary comparison, and lethal defense was only considered in the lethal attack case. Also, in all videos, the victim was human. This plan led to eight animations, shown in Table I. Animations 1-5 relate to H1, whereas 6-8 related to H2.

To implement our conceptualizations, freely available models were selected for the human, humanoid robot, and AV. Autodesk 3ds Max, Autodesk Maya, and Unreal Engine 4.27 were then used to adjust rigs and create animations.[12][13] A representative scene from each video is shown in Fig. 3, and the videos can also be seen online.[14]

D. Results

Five responses had invalid answers with wrong movie numbers, leaving 299 valid samples. Questionnaire scores, shown in Fig. 4, suggested that the idea of robot self-defense was generally accepted, as averages were all were above neutral except for one video. One-way repeated measures ANOVA was conducted for closer insight.

During an attack by a human, a difference was noted in perceived acceptance according to who was defending, in videos 1-3 ($F(2, 596) = 119.700$, $p < 0.001$). A Bonferroni post-hoc test revealed that people preferred the defender to be human rather than robot, and in the latter case, humanoid rather than an AV. Interestingly, this preference of a human over a robot defender was not observed when a robot was attacking, in videos 4 and 5: ($F(1, 298) = 0.31$, $p = .6$). However, force was perceived as more acceptable when a robot was attacking, both when the defender was a robot, comparing videos 2 and 5 ($F(1, 298) = 91.0$, $p < .001$) and when the defender was a human, comparing videos 1 and 4 ($F(1, 596) = 29.8$, $p < .001$).

Thus, hypothesis H1.1 was mostly supported: people perceived it as acceptable for a humanoid robot to use non-lethal force to protect a human, similar to, but slightly less than, use of force by a human defender (except that the slight difference was not seen when a robot was attacking). Also, hypothesis H1.2, that use of force by a humanoid robot will be perceived as slightly more acceptable than for an AV, was supported. The AV defender was perceived as least acceptable among all videos, with the only average score slightly below neutral. Thus, H1 was largely supported, that who is involved makes a difference, although both humanoid robots and humans could use force to defend.

To check hypothesis H2, robot defense behaviors were compared in the case of the human's lethal attack (video 6-8). A one-way RM ANOVA revealed a difference in perceived acceptance according to who was defending ($F(2, 596) = 238.798$, $p < .001$). A Bonferroni post-hoc test indicated that blocking was perceived as more acceptable than using non-lethal force, and that either was more acceptable than lethal force. Thus, hypothesis H2 was supported.

E. Participants' Comments

Various comments supported what was observed quantitatively, that participants felt differently when there was a robot rather than a human, and better about less force (e.g., they didn't want a robot to hurt or kill people, whose safety should be prioritized, etc.), but the comments also exposed some extra insights, which were positive, neutral, and negative:

Positive. Some participants said that it would be reassuring, convenient and great if such a robot existed, they would feel attachment to a robot that protects people, and that the robot seemed cool, righteous, and dependable, like a gentleman. Some said that all scenarios seemed acceptable, that self-defense capabilities will be necessary and important for robots in the near future, to protect the humans who make them, and that related rules will need to be considered, in line with our aim in the current work. The videos and robot movements were described as natural, smooth, realistic, easy to understand, and well-made; the idea was fun and interesting; and a nice opportunity was provided to think about how life might be like in a society in which humans and robots co-exist.

Neutral. Some participants mentioned that it was hard to assess acceptability without more information: Self-defense was deemed complex in human interactions, but even more so when a robot is added, as various factors could play a role, including the relationship between attacker, victim and

[12] autodesk.com

[13] unrealengine.com

[14] youtube.com/watch?v=F5y7dPy41p0&list=PLtGv2XOitdkQsEsPX528cmat5cnVgd15V

defender; the degree of malice in the attacker's actions; and the probability that the attacker will attack again. Another typical comment was that the robot should help *before* an attack occurs. Some participants mentioned considering risks of loss, in line with our heuristic. The videos also felt surreal to some, like a movie or game, since guns are not allowed in Japan and robots are not typical in everyday life. To allow robots to protect people, the need for a mechanism to distinguish right from wrong was also described. References were made to the movie "I Robot", regarding potential conflicts between rules governing robot behavior, and the Russo-Ukrainian War, regarding self-defense against an attacker with a firearm.

Negative. Some videos also seemed scary or shocking: At times, excessive force by powerful-looking robots was noted. Some comments also expressed concern over recognition capabilities, like if a robot might mistake play-fighting for violence, or restrain a detective trying to arrest a criminal. Other pitfalls focused on behavioral capabilities, like if a robot might not be strong enough to defend a victim against a human attacker; if malfunctioning or poor planning could hurt or kill people, including bystanders (e.g., it could be dangerous to merely push an attacker with a gun without disarming them); and if the disarming action would actually be feasible (or if there is a gap between ethically-desired behavior and reality). Some reasons were not directly connected to robots, such as an idea that returning violence for violence might not just make things worse, and a dislike for firearms. A few comments about the videos also mentioned some unnatural elements, like how pushing could appear like slapping, the robot being slow to help, and why a robot would be needed if a human can evade bullets by merely ducking.

IV. DISCUSSION

The current study presented our findings from conducting a user survey about how social robots should behave when humans in their care are threatened, in difficult cases when an altercation cannot be avoided. Quantitative analysis indicated:

- **Overall**. Our hypotheses, that the entities involved and degree of force would affect perceived acceptability, were mostly supported. Humanoid robots were accepted as defenders and preferred compared to AVs, although not quite as accepted as human defenders; also, defense against an attacking robot was more acceptable. Less force was preferred, and acceptability of self-defense was greater in life-threatening situations. These results were mostly predicted based on a single simplified heuristic, perceived risk of loss, suggesting its usefulness as a sense-making strategy in robot self-defense scenarios.
- **H1**. Our heuristic didn't clearly predict one outcome, that a significant difference in preference between human and humanoid defenders would not be seen when the attacker was a robot. Based on risk of loss, placing a human defender in the path of a robot's attack was expected to lead to lower acceptability, which was not observed. One possible explanation can be seen in a participant's comment, that they would feel more ease of mind if a human would help them; i.e., participants might have been more inclined to believe that a human defender would be capable of better defending the human victim, and been less aware of the threat posed to the human defender, who was successful in all videos.
- **H2**. It had also not been obvious to us that the humanoid robot shooting a gun would be perceived as more acceptable than the AV slapping an attacker with a car door. Participants' comments indicated that the kind of attack played a large role in their appraisals (lethal in the former case, and non-lethal in the latter video), but other secondary causes might have also contributed: People are typically familiar with cars and car-related injuries, whereas humanoid robots and guns required effort to imagine and seemed surreal/unreal to some. The use of hard objects in altercations can also be interpreted as deadly force, so any defense by a robot with a hard exterior could be perceived as potentially lethal (thus, non-lethal defense by a robot could be perceived as non-existing).

A. Future Work

This study is limited by the population (only Japanese adults), kind of study (online survey), conditions used (some combinations were not considered), and choices in how to depict characters and behaviors in the videos (the models used for the two kinds of robots examined and humans, and to represent non-lethal and lethal force via pushing and firearms). The unexpected outbreak of the Ukraine-Russian war just as the survey was being conducted might have also affected participants' perceptions of the acceptability of self-defensive against violence.

Participants' comments also provided valuable insights into the future work required to develop the necessary capabilities for self-defense robots:

- **Embodiment**. What should such a robot capable of self-defense look like? Participants commented that the robot in the videos seemed strong and scary, and that they might feel worried about dealing with such robots, but also that the robots should not be so weak that they cannot defend. A follow-up study could check if such robots could be designed to be shorter than average male height and soft (such that they can employ non-lethal force); or capable of adapting their appearance, using height in emergencies as a deterrent. Also, if lethal force from robots could be permitted, how could it be realized?–Could a robot be built with an attached firearm or knife to prevent attackers from easily stealing robots' weapons? Then too, what will happen when such robots co-exist with people? Would attackers simply destroy robots to prevent them from calling for help or filming, possibly adding new threats from ricocheting bullets, or steal robots to salvage expensive parts?
- **Recognition**. Participants mentioned the importance of reliable recognition to avoid mistakes. Various options

could be possible. In the near future, some kind of simplified command (an emergency button or sound detection), along with some kind of identification (code, fingerprint, face, or voice), could be used by a human victim to request help. Further along, autonomous capability could be developed to detect threats and attack indicators before altercations begin, and to fairly assess complex, chaotic situations once they ensue. Useful information could include who is acting as an attacker, defender, victim, or bystander (including their status, civilian or police), as well as the levels of force being used. Even the overall context might require detection; e.g., if the victim's actions have provoked the attack; if local law is being followed (e.g., a duty to retreat or stand one's ground); or if a war is going on. Attack indicators could involve combative words and tone, face expressions (glaring), gaze (looking around to see if there are no witnesses or check valuables), proxemics (advancing into personal space), posture (chest puffed, chin forward, "blading"), and hand movements (clenching fists, touching disrespectfully, concealing, and reaching into pockets).[15]
Furthermore, recognition capability will be required to detect interactional affordances; e.g., if it is possible to deescalate, distract, call for help, escape, block, use non-lethal force, or if there is no alternative but lethal force. If using a firearm, there would also need to be recognition of the backdrop and possibilities for crossfire, ricocheting bullets, or striking dangerous items in the environment such as propane tanks.

- **Behavior**. Participants mentioned the importance of avoiding unintended harm and being capable enough to prevent a human attacker, for which both high-level and low-level strategies would be required. From the high-level perspective, robots should have a clear, human-readable strategy for safely preventing, deescalating, and otherwise dealing with conflicts, including a model of potential options to consider such as a list or tree. For example, the "Five Ds" involves deflecting an attacker's weapon to avoid immediate harm, dominating (the weapon or arm or entire attacker) to avoid the weapon coming back to do harm, distracting (e.g., by causing pain to the attacker), disarming, and disabling the threat. ([16]) Also, altercations can involve multiple attackers, defenders, victims, and bystanders. Furthermore, robots should be able to follow up after an attack has been stopped: For example, in general, empathy displayed by a robot has been shown to help to reduce pain and fear in children [19]. Touches such as hugs can also be highly effective for providing social support [20]. Such work can be applied and extended. At the low-level, mechanisms for planning and conducting safe motions will be required. In dynamic, complex situations, with moving attackers, victims, bystanders, and even potentially hostages, high reliability, speed, and dexterity all seem paramount.

From a designer's point of view, there are also numerous factors to consider that were not mentioned by the participants: For example, in emergencies, trust will be important, not just in a robot, but in its developers and the wider community. Standards will be required for lawful, ethical, reliable, and valid conduct, which can be expressed with reputation or rating systems based on stars, points, or reviews. Bias can be tackled via participatory design, careful checking of data and algorithms, transparency/explainability, clear scope, continual testing, and possibility to escalate difficult cases to humans in the loop. Then also, to avoid misuse from hacking, physical connection ports should be hidden, encryption should be used, the attack surface limited as much as possible, modifications considered if common software are used (such as Robot Operating System and Operating Systems like Ubuntu), and adversarial machine learning attacks deterred (e.g., tricking visual systems, adding blind spots, and reconstructing training data) via redundancy and limited access, security audits, and clear channels for vulnerability disclosure.

Thus, the aim is to be on time with meeting people's expectations when such robot capabilities will be needed, for which additional discussion and involvement from various stakeholders will be required. By doing so, we hope that the robots of tomorrow embedded in our societies will act like rays of sunshine over what might otherwise be a darkened landscape of crime and violence, supporting a new dawn of enhanced safety and well-being for all.

References

[1] E. G. Krug, J. A. Mercy, L. L. Dahlberg, and A. B. Zwi, "The world report on violence and health," *The lancet*, vol. 360, no. 9339, pp. 1083–1088, 2002.

[2] M. A. Salichs, Á. Castro-González, E. Salichs, E. Fernández-Rodicio, M. Maroto-Gómez, J. J. Gamboa-Montero, S. Marques-Villarroya, J. C. Castillo, F. Alonso-Martín, and M. Malfaz, "Mini: a new social robot for the elderly," *International Journal of Social Robotics*, vol. 12, no. 6, pp. 1231–1249, 2020.

[3] M. M. Jung and G. D. Ludden, "What do older adults and clinicians think about traditional mobility aids and exoskeleton technology?" *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 8, no. 2, pp. 1–17, 2019.

[4] T. Halbach, T. Schulz, W. Leister, and I. Solheim, "Robot-Enhanced Language Learning for Children in Norwegian Day-Care Centers," *Multimodal Technologies and Interaction*, vol. 5, no. 12, p. 74, 2021.

[5] J. Guggemos, S. Seufert, and S. Sonderegger, "Humanoid robots in higher education: Evaluating the acceptance of Pepper in the context of an academic writing course using the UTAUT," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1864–1883, 2020.

[6] L. Cousineau and N. Miura, *Construction robots: the search for new building technology in Japan*. ASCE Publications, 1998.

[7] T. Wanebo, "Remote killing and the Fourth Amendment: Updating Constitutional law to address expanded police lethality in the robotic age," *UCLA L. Rev.*, vol. 65, p. 976, 2018.

[8] 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.

[9] M. K. Lee, S. Davidoff, J. Zimmerman, and A. Dey, "Smart homes, families, and control," in *Proceedings of the International Conference on Design and Emotion (D&E 2006*, 2006.

[15] www.scienceofpeople.com/aggressive-body-language

[16] www.facebook.com/watch/?v=773656312742151

[10] H. S. M. Lim and A. Taeihagh, "Algorithmic decision-making in avs: Understanding ethical and technical concerns for smart cities," *Sustainability*, vol. 11, no. 20, p. 5791, 2019.

[11] R. R. Galin and R. V. Meshcheryakov, "Human-robot interaction efficiency and human-robot collaboration," in *Robotics: Industry 4.0 Issues & New Intelligent Control Paradigms*. Springer, 2020, pp. 55–63.

[12] A. Gambino, J. Fox, and R. A. Ratan, "Building a stronger CASA: Extending the computers are social actors paradigm," *Human-Machine Communication*, vol. 1, no. 1, p. 5, 2020.

[13] E. K. Duarte, M. Shiomi, A. Vinel, and M. Cooney, "Robot Self-defense: Robot, don't hurt me, no more," in *2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI), Late-breaking Report (in press)*. IEEE, 2022, pp. 742–745.

[14] 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.

[15] 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.

[16] B. Rappert, "A framework for the assessment of non-lethal weapons," *Medicine, Conflict an Survival*, vol. 20, no. 1, pp. 35–54, 2004.

[17] A. Persson, M. Laaksoharju, and H. Koga, "We Mostly Think Alike: Individual Differences in Attitude Towards AI in Sweden and Japan," *The Review of Socionetwork Strategies*, vol. 15, no. 1, pp. 123–142, 2021.

[18] R. Collier, *Masculinities, crime and criminology*. Sage, 1998.

[19] M. J. Trost, G. Chrysilla, J. I. Gold, and M. Matarić, "Socially-assistive robots using empathy to reduce pain and distress during peripheral iv placement in children," *Pain Research and Management*, vol. 2020, 2020.

[20] A. E. Block, S. Christen, R. Gassert, O. Hilliges, and K. J. Kuchenbecker, "The six hug commandments: design and evaluation of a human-sized hugging robot with visual and haptic perception," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 380–388.