DOI: 10.1109/HRI53351.2022.9889590

Robot Self-defense: Robot, don't hurt me, no more*

1st Eduardo Kochenborger Duarte
School of Information Technology
Halmstad University
Halmstad, Sweden
eduardo.kochenborger-duarte@hh.se

2nd Masahiro Shiomi
Interaction Science Laboratories (ISL)
Advanced Telecommunications Research Institute International (ATR)
Kyoto, Japan
m-shiomi@atr.jp

3rd Alexey Vinel
School of Information Technology
Halmstad University
Halmstad, Sweden
alexey.vinel@hh.se

4th Martin Cooney
School of Information Technology
Halmstad University
Halmstad, Sweden
martin.daniel.cooney@gmail.com

***Abstract*—Would it be okay for a robot to hurt a human, if by doing so it could protect someone else? Such ethical questions could be vital to consider, as the market for social robots grows larger and robots become increasingly prevalent in our surroundings. Here we introduce the topic of "robot self-defense", which involves the use of force by a robot in response to violence, to protect a human in its care. To explore this topic, we conducted a preliminary analysis of the literature, as well as brainstorming sessions, which led us to formulate an idea about how people will perceive robot self-defense based on the perceived risk of loss. Additionally, we propose a study design to investigate how the general public will perceive the acceptability of a robot using self-defense techniques. As part of this, we describe some hypotheses based on the assumption that the perceived acceptability will be affected by both the entities involved in a violent situation and the amount of force that is applied. The proposed scenarios will be used in a future survey to evaluate participants' perception of a social robot using self-defense techniques under varying circumstances, toward stimulating ideation and discussion on how robots will be able to help people to live better lives.**

***Index Terms*—robot self-defense, acceptability, robot ethics, self-defense, violence**

I. Introduction

The market for social robotics has been projected to grow from USD 1.98 billion in 2020 to USD 11.24 billion in 2026, to meet needs in areas such as healthcare, education, and entertainment.[1][2] This expected increase in the prevalence of social robot technologies in our daily environments suggests the importance of exploring the practical conundrums that might emerge, in regard to what kind of robot behaviors are acceptable and what kind of expectations interacting humans should have.

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

[2]More generally, we imagine social robots providing value not just as caretakers, teaching assistants, or toys, but also as robots that we ride in, live inside, or wear: autonomous vehicles (AVs), smart homes, and wearables.

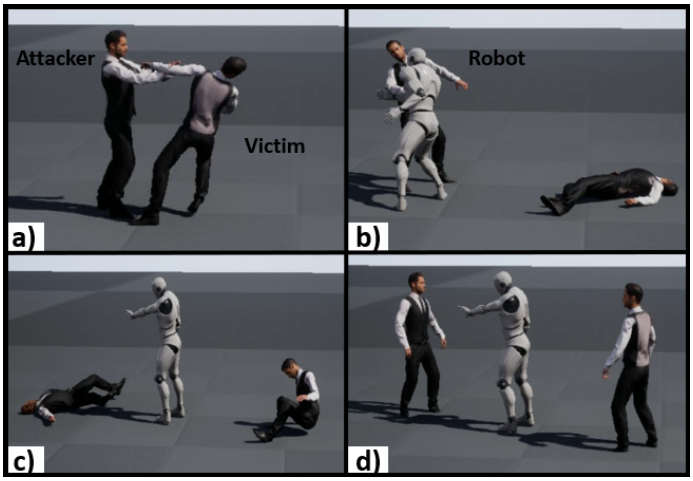

Fig. 1. Basic concept. a) An attacker (left) assaults a victim (right), b) A nearby robot comes to help, knocking down the attacker, c) The robot gets between the attacker and victim, d) Both attacker and victim get up, safe due to the robot's intervention

We propose that one fundamental conundrum that requires more discussion involves how we perceive the idea of *robot self-defense*, which does not seem to have been resolved in the literature: "Robot self-defense" here refers in a general sense to the use of physical force by a robot to counter a violent attack to some member of the robot's immediate group, including nearby humans that the robot is entrusted with caring for, as depicted in Fig. 1. On the one hand, Asimov's widely-accepted first law of robotics[3] implies that robotics researchers should not allow any kind of violence from robots [1]. Conversely, robots are often perceived as similar to humans [2], who have a "right of self-defense", and even a moral and legal "duty to rescue" in some contexts.[4]

[3]"A robot may not injure a human being or, through inaction, allow a human being to come to harm".

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

This conundrum relating to violence and robots is unlikely to be only hypothetical. Violence has been said to be a fundamental part of human existence from which we cannot escape [3], with an annual global cost of crime estimated to be in the trillions of US dollars [4]. Not only humans are the victims of violence; abusive bullying of robots has already been observed, including kicking, punching and slapping, possibly due to a perceived imbalance of power [5]. Accordingly, one study has even developed first person camera-based recognition of punching to a robot [6]. While some "home defense" robots have been developed to deter crime as moving cameras that can be used to patrol, investigate, and monitor[5], other proposals have been made for how robots could be used to enable crime, including violence and torture [7], [8]. Furthermore, robots can be built to be hard, fast, and strong, and robots have killed humans already–some accidentally (large, hard factory robots) [9], and others intentionally (military drones, even possibly "Lethal Autonomous Weapon Systems").[6]

The fundamental scenario we envision involves a robot protecting a human under its care from violence inflicted by another human, in situations that cannot be avoided by other means: For example, by fleeing with the victim, calling for help from bystanders or police, obstructing the attacker's path, pretending to have been injured by shutting down [10], simulating emotions like fear [11], or pointing out that an attack is taking place [12]. We care more about such a scenario than a dyadic one in which a robot is attacked, as replacements can be manufactured if a robot is destroyed–in line with a recent study that called into question the sagacity of hesitating to sacrifice robots to save humans [13]. Furthermore, we are most interested in the fundamental case of clear physical violence rather than verbal threats or the promise of violence, which has not been provoked by the victim, and occurs in a general "transitional space" rather than a restricted, private setting such as a home. Self-defense in such a scenario could involve the robot pushing, punching or forcibly restraining a human, or even using a "force multiplier" such as an OC spray or stun gun. Like the well-known "trolley problem", which has been extended for autonomous vehicles [14], a dilemma results from the expectation that someone will be harmed, regardless of the robot's actions: It can stand by as the person in its care is harmed, or risk harming the attacker. Will we accept a robot that could hurt us, or will we hold a robot accountable for not stepping in and helping when it could have?

Some insight into this question can be gained from sources such as fiction, legislation, and the academic literature. For example, in fiction, in the RoboCop[7] and Terminator[8] movies, among many others, robot protagonists use force to defend humans. Moreover, the television series, Black Mirror,[9] depicts various dark scenarios involving crime, violence, and technologies such as robots, for our entertainment. Yet, movies are not reality: experiences that might horrify in reality can excite us in fiction through processes like "angstlust" or "excitation transfer" [15]. Moreover, aggression and force, that we might not enjoy being the target of, can be perceived positively, according to male stereotypes, in terms of dominance, achievement, confidence, power, and mastery [16].

Legislation likewise provides some ideas, but alone seems insufficient to know what people will accept, as there is no clear consensus over all regions (e.g., there is a "duty to rescue" in most European countries, with laws also protecting good Samaritans, but for example, none in Australia).[10] This applies not only to robot self-defense, but also to the ambiguity surrounding the rules of how we should defend ourselves from robot-assisted crimes, like a drone coming to film against our wishes or an AV trespassing [17]. A further complication is the ambiguity in who is liable when an intelligent system causes harm due to algorithms not fully under the control of the creators ("res ipsa loquitur"), as when a bot purchased drugs on the Dark Web [18]–e.g., can we foretell what might happen in a chaotic, real altercation, potentially with makeshift environmental weapons?

Some ethical studies in HRI have also started to explore the role of robots in society and what robots should or shouldn't do, which has been described as Robot Ethics, Roboethics, Machine Ethics, Artificial Moral Agents, or Friendly AI–for example, considering relevant concepts such as "moral agents" vs. "moral patients", "mind perception", "moral dyads", and the "harm-made mind" [19]. In this area, one common proposal is that robots will be particularly suited for dangerous tasks [20], which could include protecting humans. However, it has been noted that most current robotics research has focused on utilitarian needs rather than on values and morals, offering an explanation for why little seems to have been done yet on this topic of robot self-defense [21]. The same study furthermore proposed how a seemingly negative combination of destruction and robots could actually facilitate positive experiences involving creation, catharsis, and emotional support: similarly, could a robot protecting a bullied child empower them, or even motivate others to also rise up to protect the vulnerable?

Additional questions arise when considering the numerous individual factors that could affect how robot self-defense is perceived: The kind of robot embodiment is likely to have an effect; e.g., humanoid appearance and size have been observed to affect perceived intelligence, neuroticism, and conscientiousness, which could play a role in how self-defense is perceived [22]. Moreover, various degrees of force exist, which could be reasonable and matched to the threat, ineffective, or unjust and excessive [23]. Also, various other variables such as culture could play a difference, although

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

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

[7]en.wikipedia.org/wiki/RoboCop_(franchise)

[8]en.wikipedia.org/wiki/Terminator_(franchise)

[9]en.wikipedia.org/wiki/Black_Mirror

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

some recent studies have argued that there are less differences due to culture between some countries, such as Sweden and Japan, than has been previously thought [24].

Thus, the literature offered insight, but did not indicate whether people would accept robot self-defense or not.

II. Methodology

Our analysis of the literature suggested that there seem to be some unanswered questions related to the acceptability of robot self-defense.

To begin to investigate this topic, brainstorming was conducted, both offline and online. First, our group prepared a shared document containing early ideas, insights, and questions. Then we held an online workshop meeting around these ideas (seeded brainstorming). At the start of the workshop, we followed an "excursion"-like approach to seek to stimulate creativity and flexible thinking; this involved listening to relaxing music while viewing some randomly selected images from the Oasis emotional database, which included both positive and negative scenes of people and machines [25]. (A Python program was used to automatically select photos, which were different per person.)

This approach was loosely based on the idea of the New Metaphors Toolkit, which has been used for HRI to develop new ideas [26]. Furthermore, Dotmocracy was used to identify topics we were interested in investigating, shaping the emphasis.[11] In particular, two kinds of social robot that seemed interesting included a standard humanoid robot, as well as a large, mechanical robot (an AV). Based on this, hypotheses were formulated. Thereafter, we used the "How might we" approach[12], to move from our insights and hypotheses to how the animations could be specified. A collection of freely available models were gathered and the most suitable were selected. For adjusting the models (e.g., the rigs), Autodesk 3ds Max was used, and for developing new animations, Autodesk Maya was used.[13] The models were then imported in Unreal Engine 4.27[14], where a new sequence was created for each scenario using a combination of newly created and sample animations.

III. Preliminary Results

Based on our discussions, we believe that:

1 *Who* is involved in an attack will affect how acceptable self-defense is. As noted, robots are often perceived like humans (somewhere between humans and objects), so it will likely be perceived as acceptable also for a robot to defend against a human with non-lethal force. However, moral judgements can involve utilitarian (hedonic) appraisals, in line with the idea that actions should seek to lead to the greatest happiness for all involved [27]; or in this case, that the perceived *risk of loss* should be minimized. Specifically, it will be perceived as more acceptable for a robot to defend itself if the potential loss due to the attack is large and the potential loss due to defending is low, also from the perspective that human life cannot be replaced, whereas robots can be fixed. Thus, a robot defending against a human will be perceived as less acceptable than a human defending against a human. In the former case, the robot could harm a human by resisting, but no human will be harmed if the robot does not defend. In the latter case, both defending or not defending could lead to harm. Furthermore, the most acceptable defense 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. (For example, age or disability could affect how strong a person appears.) A humanoid robot looks more human than a mechanical robot such as an AV, so it might be perceived as more acceptable for such a robot to defend itself; however, an AV that does not defend itself could be taken over and used as a weapon of deadly force.

2 *What* behavior is conducted will affect how acceptable self-defense is. As above, participants' perceptions will be affected by the potential loss: Non-lethal force will be more acceptable than lethal force, and less acceptable than blocking. (Running would be most acceptable in a dyadic case, but in a triadic or tetradic case could mean leaving a human vulnerable to attack and thus is not considered here to be a self-defense behavior. Likewise, calling for help does not prevent an attack in the envisioned scenario.) A humanoid robot typically is small and weak to avoid harming or alarming people, whereas an AV is large and easily capable of killing a human; thus, it will be perceived as more acceptable for a humanoid robot to defend itself.

This ideation led to two main hypotheses: *H1 Embodiment. H1.1* People will perceive it as acceptable for a robot to use non-lethal force to protect a human, similar to, but slightly less than, use of force by a human defender. *H1.2* Use of force by a humanoid robot will be perceived as slightly more acceptable than for an AV. *H2 Behavior.* The less force is used by a robot, the more acceptable.

To test the hypotheses, the plan is to use a questionnaire in two parts: In the first part, participants will watch a single animation and describe how much they agree with the statement, "The defender's behavior is acceptable". In the second part, participants will compare two videos and state which is more acceptable.

This plan led to eight animations: *Animation 1:* Human defends human against human (non-lethal defense). *Animation 2:* Robot defends human against human (humanoid, non-lethal defense). *Animation 3:* Robot defends human against human (AV, non-lethal defense). *Animation 4:* Human defends human against robot (humanoid, non-lethal defense). *Animation 5:* Robot defends human against robot (humanoid, non-lethal defense). *Animation 6:* Robot defends human against human (humanoid, lethal defense, lethal attack). *Animation 7:* Robot defends human against human (humanoid, non-lethal defense, lethal attack). *Animation 8:* Robot defends human against human (humanoid, block defense, lethal attack). Animations 1-5 relate to H1, whereas 6-8 related to H2.

[11] dotmocracy.org/what_is

[12] odellkeller.com/the-how-might-we-method/

[13] autodesk.com

[14] unrealengine.com

IV. Discussion

Our preliminary results include the following:

- **Robot Self-defense**. We pointed out a new idea regarding a problem that we think could be important for social robots in the near future, how robots should be expected to deal with violence.
- **Theory**. We explored what insights can be derived from the literature, and proposed a concept of expected loss that might allow us to guess how people will perceive robot self-defense.
- **Next steps**. We furthermore developed a plan for a user study to test our hypotheses, including eight videos.

In general, our ideation so far suggests that a blind eye has been turned up until now to some of the darker questions in HRI, but can we afford to continue down this road without discussing such challenges?–and if so, what might be the cost?

References

[1] 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.

[2] 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.

[3] T. Hobbes, *Leviathan Or the Matter Form and Power of a Commonwealth, Ecclesiastical and Civil*. London:[sn], 1886, vol. 21.

[4] M. DeLisi, "Measuring the cost of crime," *The handbook of measurement issues in criminology and criminal justice*, pp. 416–33, 2016.

[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] L. Xia, I. Gori, J. K. Aggarwal, and M. S. Ryoo, "Robot-centric Activity Recognition from First-person RGB-D Videos," in *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 357–364.

[7] N. Sharkey, M. Goodman, and N. Ross, "The coming robot crime wave," *Computer*, vol. 43, no. 8, pp. 115–116, 2010.

[8] T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi, "Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions," *Science and engineering ethics*, vol. 26, no. 1, pp. 89–120, 2020.

[9] L. A. Kirschgens, I. Z. Ugarte, E. G. Uriarte, A. M. Rosas, and V. M. Vilches, "Robot hazards: from safety to security," *arXiv preprint arXiv:1806.06681*, 2018.

[10] 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.

[11] G. Hoffman, O. Zuckerman, G. Hirschberger, M. Luria, and T. Shani-Sherman, "Design and evaluation of a peripheral robotic conversation companion," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 3–10.

[12] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using robots to moderate team conflict: the case of repairing violations," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 229–236.

[13] K. Mamak, "Whether to save a robot or a human: on the ethical and legal limits of protections for robots," *Frontiers in Robotics and AI*, vol. 8, 2021.

[14] 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.

[15] D. Zillmann, "Excitation transfer theory," *The international encyclopedia of communication*, 2008.

[16] J. T. Wood, "Gendered media: The influence of media on views of gender," *Gendered lives: Communication, gender, and culture*, vol. 9, pp. 231–244, 1994.

[17] A. M. Froomkin and P. Z. Colangelo, "Self-defense against robots and drones," *Conn. L. Rev.*, vol. 48, p. 1, 2015.

[18] R. Calo, "Robotics and the lessons of cyberlaw," *Calif. L. Rev.*, vol. 103, p. 513, 2015.

[19] 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.

[20] L. Takayama, W. Ju, and C. Nass, "Beyond dirty, dangerous and dull: what everyday people think robots should do," in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2008, pp. 25–32.

[21] 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.

[22] M. L. Walters, K. L. Koay, D. S. Syrdal, K. Dautenhahn, and R. Te Boekhorst, "Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials," *Procs of New Frontiers in Human-Robot Interaction*, 2009.

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

[24] 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.

[25] B. Kurdi, S. Lozano, and M. R. Banaji, "Introducing the open affective standardized image set (OASIS)," *Behavior research methods*, vol. 49, no. 2, pp. 457–470, 2017.

[26] P. Alves-Oliveira, M. Luce Lupetti, M. Luria, D. Löffler, M. Gamboa, L. Albaugh, W. Kamino, A. Ostrowski, D. Puljiz, P. Reynolds-Cuéllar *et al.*, "Collection of Metaphors for Human-Robot Interaction," *Proceedings of the 2021 ACM:*, 2021.

[27] J. Bentham, "An Introduction to the Principles of Morals and Legislation (1789)," *H Burns and HLA Hart, London*, 1970.