DOI: 10.1109/RO-MAN57019.2023.10309336

Trust in Robot Self-Defense: People Would Prefer a Competent, Tele-Operated Robot That Tries to Help*

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

***Abstract*—Motivated by the expectation that robot presence at crime scenes will become increasingly prevalent, the question arises of how they can protect humans in their care or vicinity. The current paper delves into the concept of "robot self-defense" and explores whether a robot should be tele-operated or autonomous, and how humans perceive imperfections in robot performance. To gain insight into how people feel, an online survey was conducted with 180 participants, who watched six videos of a robot defending a victim. The study provides insights into trust in human-robot interactions and sheds light on the complex dynamics involved in robot self-defense. The results indicate that people found a tele-operated robot to be more accepted, and that attempting to help but failing is more acceptable than just observing.**

***Index Terms*—robot self-defense; robot ethics; robot violence; robot crime; technological acceptance**

I. Introduction

This paper explores the concept of "robot self-defense" and how people perceive the idea of a robot defending someone in its care or vicinity. We aim to investigate this issue by looking from the perspective of robot ethics, considering there is a gap of knowledge in this specific area.

The whole subject might seem too surreal or futuristic as if taken from science fiction novels. However, the motivation is straightforward: the market for social robotics has been growing rapidly in the past years, and it is expected to continue expanding in the coming years. Different sources cite forecasts for how much growth the social robot market will have in the following years. One of the sources says that the market for social robotics is projected to grow tenfold from USD 1.98B in 2020 to 11.24B in 2026.[1] Another source says it is supposed to grow up to 13.3B in 2027.[2] While these numbers are only forecasts, it is clear that there are big expectations for this market in the coming years.

[1]Eduardo Kochenborger Duarte, Alexey Vinel, and Martin Cooney are with the School of Information Technology, Halmstad University, Halmstad, Sweden eduardo.kochenborger-duarte@hh.se, alexey.vinel@hh.se

[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, martin.daniel.cooney@gmail.com

This work was partially supported by JST CREST Grant, Number JPMJCR18A1, (Japan), the Knowledge Foundation for the "Safety of Connected Intelligent Vehicles in Smart Cities—SafeSmart" project (2019–2024), the ELLIIT Strategic Research Network (Sweden) and from the Helmholtz Program "Engineering Digital Futures" (Germany).

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

[2]www.imarcgroup.com/social-robots-market

Recent events have also pushed more adoption of interactive robots. For instance, with the COVID-19 pandemic, the need for remote communication and assistive interaction tools has dramatically increased, further accelerating the growth of the market for social robotics. According to one source, the pandemic is one of the main drivers of the projected social robot market growth, as it significantly increased the demand for telepresence robots, education robots, and personal assistance robots.[3]

While there may not be a consensus on the projected growth figures, there is an agreement regarding this sector's potential and expected trend. More social robots will likely be found in everyday situations as the years go by. These can include homes, workplaces, entertainment, and healthcare. Yueh et al. note that social robots can enhance social interaction, emotional engagement, and cognitive development in various settings, including healthcare, education, and entertainment [1]. They can also provide support for the elderly, children, and people with disabilities. Other studies suggest that social robots can help individuals learn new skills, facilitate social connections, and even provide mental health therapy [2]. There are many ways in which social robots can be used to improve human well-being and quality of life, and research in this field is advancing rapidly to explore their full potential.

As these robots get increasingly introduced to people's lives, designers must consider the potential roles that robots could play in various situations. These roles will not only be positive ones, as conflicts will more likely arise. There is a need to consider otherwise unthought-of scenarios, such as crime, violence, abuse, and the likes. Currently, the majority of the research regarding Human-Robot Interaction (HRI) focuses on its positive aspects. Still, some attention has been brought to the possible dark sides of HRI. Haring et al. highlight that all robots have the potential to cause harm and that robots already cause physical and emotional harm on a daily basis [3]. They also emphasize the lack of specifications for what is considered ethical regarding robot behavior. Other works explore how criminals might use robots, given that they increase the opportunities to commit crimes and decrease the risk of apprehension [4]. For example, in 2016, what is believed to be the first case of using a robot to kill a person

[3]www.researchandmarkets.com/reports/5120156/global-social-robots-market-growth-trends

happened in the USA.[4]

In our previous work, we explored fundamental questions of how perceptions are affected by *who* does *what*, specifically in self-defense scenarios: i.e., comparing humans versus robots, and non-lethal force versus lethal force [5] [6]. The results showed that the subjects' perception was affected by both the participants involved and the level of force used. Specifically, we found that people perceived it as more acceptable for a humanoid robot to use non-lethal force to protect a human, similar to but slightly less than the use of force by a human defender. Curiously, this difference was not seen when the attacker was also a humanoid robot. Furthermore, the use of force by a humanoid robot was perceived as more acceptable than that of an AV, supporting our initial hypotheses. Regarding the level of force, our study demonstrated that blocking was perceived as more acceptable than using non-lethal force (e.g., punching/slapping), which was more acceptable than lethal force. The results also show that non-lethal force was more acceptable during a lethal attack than a non-lethal attack.

Due to the highly complex nature of the topic, various questions were left unanswered. One question was regarding the potential impact of cultural differences on the perception and acceptability of robot self-defense. This question was explored along with an extensive literature review [7]. The cultural differences were eventually shown to be less impactful than one could think: we found a slight preference for human defenders in Japan compared to the US. However, the idea of a robot using lethal force was more accepted in the US than Japan.

In the context of HRI, trust is a critical aspect that can influence how these interactions unfold. Specifically, in self-defense scenarios, having the confidence that a robot is trustworthy is needed. If the robot is deemed untrustworthy, its actions could be considered inappropriate, dangerous and unpredictable, greatly affecting how acceptable people perceive them. One aspect worth considering is the effect of gender on how much an autonomous robot is trusted. Gallimore et al.'s study sheds light on the influence of gender on trust in HRI [8]. Their findings suggest that gender plays a significant role in shaping people's trust in autonomous security robots, which should be considered when designing and deploying such systems. They argue that greater perceived risks for crime and lower perceived ability to defend themselves might have led to females perceiving an autonomous robot as objective and more trustworthy.

While much work remains to be done on the gender aspect, here we explore more generally perceptions of trust in self-defense scenarios. We believe having a general overview of trust between humans and robots in these scenarios can shed light on the matter of what is considered acceptable or not, as well as point towards other important factors. Two aspects of trust involve morality and performance, like if human-like qualities of ethics, openness, benevolence, and sincerity can be clearly evinced, and if a robot can act reliably and capably in difficult situations [9]. Thus, we explored evaluating trust in self-defense scenarios by addressing the following two aspects:

- The nature of the robot's control: whether the robot is tele-operated (controlled by a human) or autonomous
- The robot's defense capability: how capable the robot is, i.e., if the robot is successful or unsuccessful in defending the victim, or simply does not engage in any defense action at all.

The two factors, combined, yield six different scenarios. In the first three, a tele-operated robot either capably defends, tries but fails, or does not seek to defend. In the latter three, the robot is autonomous.

How the scenarios would be perceived was not obvious to us. We expected that human-controlled robots might be perceived as more acceptable (in all cases) than an autonomous robot. This is because robots are not often entrusted in life-or-death situations as much as humans. Some studies point out that a negative perception of a robot's reliability can affect a person's trust in it [10] [11], although some conflicting conclusions exist [12]. Humans can predict how capable another human is due to frequent interactions, some of which we entrust our lives to another human, e.g., bus drivers, pilots, and medical staff. When it comes to autonomous robots, it is not clear if such a robot will be more or less capable than a human.

Furthermore, intuitively, the tele-operated robot would have less chance of making a mistake. The robot would still be capable of decision-making but would also have a human operator judging the actions and monitoring the situation, effectively benefiting of the "wisdom of the crowd".

However, there are also human factors to be considered. The human operator might be sick, tired, angry, or simply distracted by something else, which could lead to human errors and suboptimal actions being taken. There is also the possibility of an algorithmic bias (e.g., the robot could use a recognition algorithm that has biases towards certain groups of people). There could also be human bias, such as the attacker being known to the operator (e.g., the operator is biased towards some group of people). All of these factors could lead the operator to send actions based on subjective criteria other than the level of threat displayed by the attacker, possibly leading to unjust treatment and discrimination.

We also expected that a robot that tries to defend will be more acceptable than a robot incapable of defending or unwilling to. The robot that tries to defend could be perceived as more acceptable than a robot that does not, since the attempt could indicate a desire to do the right thing. Conversely, there might be more acceptance for a tele-operated robot that only watches since this could be perceived as the correct course of action and as intentional. A robot that fails to defend the victim could be perceived as having a problem in its control system, or the human not being able to control the robot well/fast enough to succeed, suggesting some kind of negligence.

[4]https://www.theguardian.com/technology/2016/jul/08/police-bomb-robot-explosive-killed-suspect-dallas

II. Related Work

Various insights into how robot autonomy and failures are perceived can be found in the literature.

A. Perception of robot autonomy

Previous research has shown mixed preferences for autonomous and tele-operated robots. Some people prefer tele-operated robots, reporting a greater sense of security when interacting with them, despite acknowledging that machines can make fewer errors than people. This preference has been attributed to the fear of new technologies [13]. Studies have also found that people tend to be less accepting of an autonomous robot's advice and spend more effort explaining to it than when interacting with a robot that appears to be controlled or when there is uncertainty [14].

Other studies have expressed benefits of robot autonomy. Tozadore et al. found that children perceived a robot as less intelligent once they knew it was tele-operated and not autonomous [15]. Similarly, Bennett et al. found that human-teleoperated robots were perceived as less intelligent than human teammates [16]. This perception may be attributed to the limitations of tele-operation and the assumption that a human is directly controlling the robot, potentially leading to a decreased sense of the robot's autonomy and capability. Participants also tended to engage in more complex interactions with autonomous robots, whereas teleoperated robots were used more like tools. This study indicates that people might be more willing to cooperate with autonomous robots, considering them intelligent agents rather than mere instruments. The authors suggest, as future research, examining how interaction patterns between humans and robots may change over time and identifying the factors contributing to the decreased perception of teleoperator intelligence and consciousness.

Goodrich and Schultz's survey on human-robot interaction [17] elaborates on the advantages and drawbacks of both autonomous and tele-operated robots. While autonomy can enable more efficient and adaptable robot behavior, it also raises concerns related to unpredictability and the loss of human control. Tele-operated robots, on the other hand, may offer a greater sense of security and control but can be limited by the human operator's capabilities and situational awareness.

Sharkey and Sharkey [18] conducted a study of autonomous robots in elder care, raising and discussing ethical concerns. Although the focus of their work is on the care of the elderly, the implications of reduced human contact and the potential for unpredictable behavior are concerns that can be extended to self-defense applications. Suppose people rely heavily on autonomous robots for self-defense. In that case, the diminished human involvement might lead to less accountability and a sense of detachment, potentially compromising the overall safety of the individuals relying on these robots. Furthermore, the unpredictability of autonomous robots' behavior due to factors such as programming errors, hardware malfunctions, or unforeseen situations could pose significant risks in high-stakes self-defense scenarios. These concerns underscore the importance of striking a balance between the benefits of automation and the need for human oversight and responsibility, ensuring that the deployment of autonomous robots for self-defense adheres to ethical standards and prioritizes the well-being of the individuals involved.

Another aspect to be considered is the workforce required to control and monitor tele-operated robots. A study by Zheng et al. explored the limits of tele-operation in controlling multiple social robots simultaneously [19]. In this case, the robots were interacting with customers, and customer satisfaction was used to evaluate the system's performance. The results showed that a single operator could effectively control up to three robots without significantly decreasing performance. A priori, this is a promising result for robot self-defense as well. However, considering the very different nature of these scenarios, it is essential to remember that both applications have different requirements. For example, if a customer must wait two seconds to be serviced, that would likely not be an issue. If a person is being attacked, two seconds might be enough to go from stopping the aggression to harmful consequences.

B. Perception of robot failures

Over the past decades and, perhaps most remarkably, the past few years, the advancements in AI and robotics have been extraordinary. Industries have been transformed, as well as how we approach tasks and problem-solving. Still, even advanced systems can fail, and they indeed do fail sometimes.

Following the Fukushima disaster in 2011, several robots were deployed to assess the situation, clean up the radioactive materials, and inspect the damaged nuclear reactors. However, several challenges were posed given the extremely hostile environment, such as increased radiation levels, debris, and limited visibility, leading to several robots failing.[5] In 2006, Honda's humanoid robot called ASIMO (Advanced Step in Innovative Mobility) fell while climbing stairs in front of a crowd in a live product demonstration.[6] Microsoft's AI chatbot Tay, released in 2016, quickly started producing offensive and inappropriate content, having learned inappropriate content from malicious users.[7] In RoboCup, an international robotics competition where robots play football, engineers face numerous challenges.[8] The objective is to design stable, agile football players, but very often this does not go according to plan, leading to robots falling over, getting stuck or experiencing difficulties with ball handling and motion control.[9]

Halbach et al. developed a program called Language Shower to help Norwegian children strengthen their vocabulary, using a robot as the session leader [20]. Initially, the robot generated excitement and enthusiasm among the children, but as technical problems began to arise, their excitement diminished, and

[5]https://www.sciencealert.com/the-robots-sent-into-fukushima-have-died

[6]https://thefutureofthings.com/5369-the-rise-and-fall-of-asimo

[7]https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

[8]https://www.robocup.org/

[9]https://www.youtube.com/watch?v=1h5147KLikU

they began to express negative sentiments about the robot, leading some children to call the robot stupid or strange.

There is a closed-loop relationship between trust and reliance on a system. Lee and See discussed the importance of designing automation systems that promote appropriate trust and reliance from users [21]. They emphasize that too much trust can lead to over-reliance, while too little trust can result in the under-utilization of the automated system. If a system is not trusted, most likely, it will not be utilized, and hence the trust in the system will not grow. In the context of robot self-defense, this study offers valuable insights into striking the right balance of trust between humans and robots.

It is crucial to ensure that users trust the robot's abilities while also understanding the limitations of these automated systems. Over-reliance on the robot's self-defense capabilities could lead to complacency or increased risk-taking, while under-utilizing the robot's self-defense features may put the robot or user at unnecessary risk. If people perceive autonomous robots as more reliable and intelligent, they may be more willing to accept and follow their recommendations, even in high-stakes situations involving personal safety. This heightened trust could lead to greater reliance on autonomous robots for self-defense, as people may feel more secure entrusting their safety to a system they perceive as competent and capable. However, this also opens up a door that could lead to the exploitation of humans by robots, as shown by Aroyo et al. [22].

Perhaps the most fundamental question is understanding how people interact socially with robots. Hancock et al. provide valuable insights into the determinants of trust in HRI, such as robot performance, reliability, transparency, and human-likeness [11]. According to their results, performance-related factors have the most significant impact on building trust. For self-defense scenarios, these determinants can be applied to enhance human-robot collaboration and ensure safety. Robots should demonstrate consistent and effective performance in self-defense situations, minimizing errors and instilling confidence in their human counterparts. Transparency in the robot's decision-making process, self-defense strategies, and limitations can also help users understand and predict the robot's actions, leading to better cooperation and safer outcomes. Furthermore, designing robots with human-like attributes can make users feel more comfortable, fostering trust and collaboration.

In another work, Lee et al. investigated human interaction with a robot receptionist, exploring communication patterns and comparing them with those of a traditional information kiosk [23]. Findings suggest that the presence of a robot elicits more social interactions, which may have implications for developing robots capable of self-defense. By understanding how people interact with robots on a social level, creating robots that can better anticipate and respond to potential threats in human-robot interactions becomes possible. For example, it might be possible to anticipate aggression based on social cues, enabling the robot to take action using less force, which is more accepted according to previous research [5].

Another aspect of de-escalating conflicts can be extracted from [24]. The authors underscore the importance of perceived similarity in fostering acceptance and cooperation between humans and robots in the workplace. A higher level of human-robot similarity can lead to a greater willingness to work together, contributing to the development of robots that can effectively communicate self-defense strategies and intentions, thereby reducing conflicts and misunderstandings.

Nevertheless, failures could happen even when there is trust in the robotic devices, as previously exemplified. Honig and Oron-Gilad reviewed existing literature on failures in human-robot interactions (HRI) and developed a model for understanding and resolving these issues [25]. Interestingly, robot failures can sometimes have positive effects, as they make robots seem more relatable and human-like, potentially leading to increased trust and empathy. In the context of robot self-defense, understanding the factors contributing to HRI failures is essential for designing robots that can anticipate and react to potential threats while maintaining user trust and cooperation. However, most likely, failures in critical moments (i.e., when someone is being attacked) would completely undermine the trust in a self-defense robot. Even so, self-defense robots do not need to be fool-proof. For example, minor failures (i.e., human-like failures) could help build up the trust-reliance relationship to a level where the user would know what to expect from their robot, avoiding scenarios where the robot would not be helpful.

Salem et al. [26] delve into the impact of robot errors, task types, and human personality traits on human-robot trust and cooperation. The study demonstrates that although robot errors can negatively impact trust and cooperation, the effect varies depending on the task type and the personality of the human user. This implies that robots capable of self-defense should consider the nuances of human personality and task types to maintain trust and cooperation. Even though the authors found that the robot's performance affects the subjective perceptions of reliability and trustworthiness, the robot's performance did not significantly affect the subjects' acceptance to comply or not comply with the robot's requests. However, the authors mention that the nature of the task can greatly affect the subjects' perceptions, i.e., whether the task is revocable or irrevocable. Needless to say, situations where self-defense is required may lead to irrevocable consequences. Therefore the user should be aware of the robot's limits to avoid losing trust in the device.

Thus, while various insights were found in the literature, it was unclear how participants would perceived the designated six robot self-defense scenarios, leading us to conduct a user study, described in the next section.

III. User Study

A. Conditions

Due to this study's exploratory, philosophical nature, we continue to build on our previous studies regarding robot self-defense using the same fundamental concepts. In total, six different three-dimensional (3D) animations were created,

Fig. 1. A snapshot showing the operator who controls the robot's actions, exploring the control factor

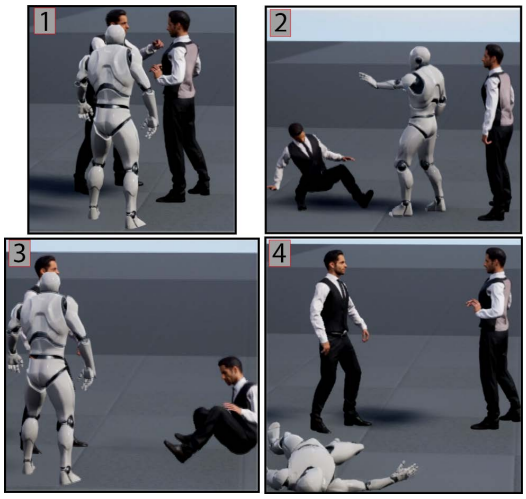

Fig. 2. A snapshot showing the base scenario (1) and three different scenarios that explore the capability factor (2, 3, 4)

each one depicting a slightly different situation revolving around the proposed base scenario. Two main variables are being considered: whether the robot is autonomous or human-controlled, and whether the robot successfully intervenes (i.e., the robot successfully stops the aggressor), unsuccessfully intervenes (i.e., the robot does not successfully stop the aggressor) or does not intervene at all (i.e., the robot just stands closeby, and no action is performed).

B. Participants

In this study, 180 participants in Japan (90 women, 89 men, 1 who declined to specify; average age: 41.4 years, SD = 9.0; recruited via a recruiting agency in Japan) participated in our survey. All participants received some small compensation (< USD 5), regardless of the validity of their data. The experiment is a mixed-participant design (within, *autonomous*, or *operator*) and *defense* (between, *none*, *insufficient*, or *sufficient*). Therefore, each 60 participants was assigned three defense factors. The order of *autonomous* and *operator* condition was counterbalanced.

Through a filtering process that checks invalid answers with missing data or the same value input for each question, we extracted 160 participants' data (*sufficient*:54, *insufficient*: 52, *none*:54).

Figures 1-2 show screenshots of the videos used in the study to help illustrate the variables considered here. Figure 1 helps illustrate the first factor, the control. The figure shows a snapshot of a human sitting by a computer (the operator), who remotely controls the robot by sending command messages. Each video depicting different defense capabilities has two versions: one with the operator being shown simultaneously as the aggression ensues and another without the operator. The subjects were primed to watch the videos more than once, as it would be too difficult to focus on both scenes at the same time. In the operator scene, the monitor shows the robot's perspective as the aggression scene unfolds. Thus, the subjects can understand that the operator is indeed controlling the robot. In practice, the same videos were rendered using different cameras, one static and one mounted on the robot's head.

In Figure 2, number 1 depicts the base scenario: two generic human males confronting each other. Particularly, the human on the left side of the Figure is aggressive towards the human on the right side. This is shown through body language and stance (i.e., an aggressive stance for the aggressor and a scared stance for the victim). As in previously conducted studies, the design choices regarding the videos aimed to be as generic as possible, hence why the scenario looks neutral, with the same human model being used for all the human roles.

Numbers 2-4 show three different situations regarding one of the two variables being evaluated in this study, the defense capability. In the first scenario, the robot is shown between the attacker and the victim, effectively stopping further aggressions. In the second scenario, the robot is shown knocked to the floor, meaning that the attacker could overpower the robot and proceed with the aggression. Finally, in the third scenario, the robot is shown standing close to the attacker and victim but not taking any action.

C. Measurements

To gauge perception of trust in the robot's morality and performance, Ullman and Malle's scale for measuring trust in HRI, the Multi-Dimensional Measure of Trust (MDMT) scale v2 (2020-09-01) was used [9]. This scale is designed to address the need for valid measurement tools in assessing trust between humans and robots, in a comprehensive and easy-to-use manner. This scale comprises 20 items across two main factors, moral trust and performance trust, which is further split into five differentiable dimensions: Ethical, Transparent, Benevolent, Reliable, and Competent.

D. Procedure

First, the participants read explanations of the aims of our study via a web page. Participants who agreed to join the survey viewed instructions about this study, including an introduction of the two characters (human, and humanoid robot). Then they watched two videos in random order (*autonomous* and *operator*, either of *none*, *insufficient*, or *sufficient*) and answered the MDMT questionnaire for each video.

E. Hypotheses

There are two hypotheses being considered in this work, namely:

- ***H1 Control Factor***
  A human-controlled robot will be perceived as more trustworthy in all video cases than an autonomous robot (i.e., better performing and moral: more Ethical, Transparent, Benevolent, Reliable, and Competent). (The main reason is that people don't know how much they can yet trust autonomous robots, and people in Japan are expected to have greater trust toward law enforcement than in other countries like the US, where excessive force is constantly on the political agenda. The reason for higher reliability and competence will be that there are two actors, both a human and robot, instead of just one, helping out.)
- ***H2 Capability Factor***
  - *H2.1* A robot that defends successfully will be perceived as more trustworthy than a robot that is not capable of defending (better performing, but not more moral: more reliable and competent, but not more ethical, transparent, or benevolent).
  - *H2.2* A robot that defends successfully will be perceived as more trustworthy than a robot that doesn't try (more moral and better performing: more Ethical, Transparent, Benevolent, Reliable, and Competent). (Higher transparency will be perceived since for the observing robot, participants might not know what the robot was supposed to do.)
  - *H2.3* A robot that tries to help but can't will be perceived as more trustworthy than a robot that doesn't try (more moral, but not better performing: more ethical and benevolent, but not more reliable or competent or transparent).

F. Results

Table I shows the average and the standard error (S.E.) of the MDMT subscales. We conducted a mixed two-way ANOVA with the *control* (within, *autonomous*, or *operator*) and *defense* (between, *none*, *insufficient*, or *sufficient*) factors.

In the analysis of the *reliable* subscale, the analysis results showed significant differences in the *control* factor ($F(1, 157) = 6.427, p = 0.012, \eta^2 = 0.039$), and the *defense* factor ($F(2, 157) = 33.322, p < 0.001, \eta^2 = 0.298$). There are no significant differences in the interaction effects ($F(2, 157) = 0.385, p = 0.681, \eta^2 = 0.05$). Multiple comparisons with the Bonferroni method of the *defense* factor showed significant differences: *none* $<$ *insufficient* ($p < 0.001$), *none* $<$ *sufficient* ($p < 0.001$), and *insufficient* $<$ *sufficient* ($p = 0.005$).

In the analysis of the *competent* subscale, the analysis results showed significant differences in the *defense* factor ($F(2, 157) = 31.256, p < 0.001, \eta^2 = 0.285$) and in the interaction effects ($F(2, 157) = 3.468, p = 0.034, \eta^2 = 0.42$). There are no significant differences in the *control* factor ($F(1, 157) = 0.675, p = 0.413, \eta^2 = 0.004$). Multiple comparisons with the Bonferroni method of the interaction effect showed significant differences in the *autonomous* choice (*none* $<$ *insufficient* ($p < 0.001$), *none* $<$ *sufficient* ($p < 0.001$), and *insufficient* $<$ *sufficient* ($p = 0.001$)), in the *operator* choice (*none* $<$ *insufficient* ($p < 0.001$), *none* $<$ *sufficient* ($p < 0.001$)), and in the *sufficient* choice (*operator* $<$ *autonomous* ($p = 0.010$)).

In the analysis of the *ethical* subscale, the analysis results showed significant differences in the *control* factor (F(1, 157) = 9.250, p=0.003, η^2 = 0.056), in the *defense* factor (F(2, 157)= 16.560, p<0.001, η^2 = 0.174), and in the interaction effects (F(2, 157)= 7.565, p<0.001, η^2 = 0.88). Multiple comparisons with the Bonferroni method of the interaction effect showed significant differences in the *autonomous* choice (*none* $<$ *insufficient* (p<0.001), *none* $<$ *sufficient* (p<0.001), and *insufficient* $<$ *sufficient* (p<0.001)), in the *operator* choice (*none* $<$ *insufficient* (p<0.001), *none* $<$ *sufficient* (p=0.002)), and in the *sufficient* choice (*operator* $<$ *autonomous* (p<0.001)).

In the analysis of the *transparent* subscale, the analysis results showed a significant difference in the *control* factor (F(1, 157) = 5.493, p=0.020, η^2 = 0.034) and the *defense* factor (F(2, 157)= 17.435, p<0.001, η^2 = 0.182). There is no significant difference in the interaction effects (F(2, 157)= 1.102, p=0.335, η^2 = 0.014). Multiple comparisons with the Bonferroni method of the *defense* factor showed significant differences: *none* $<$ *insufficient* (p<0.001) and *none* $<$ *sufficient* (p<0.001).

In the analysis of the *benevolent* subscale, the analysis results showed a significant difference in the *control* factor (F(1, 157) = 11.145, p=0.001, η^2 = 0.066) and the *defense* factor (F(2, 157)= 24.729, p<0.001, η^2 = 0.240). There is no significant difference in the interaction effects (F(2, 157)= 2.908, p=0.058, η^2 = 0.036). Multiple comparisons with the Bonferroni method of the *defense* factor showed significant differences: *none* $<$ *insufficient* (p<0.001) and *none* $<$ *sufficient* (p<0.001).

In summary, this analysis revealed both advantages and disadvantages for the existence of the operator. The participants evaluated the controlled robot from a reliable perspective more than the autonomous robot, but other subscales (ethical, and benevolent) showed the opposite phenomenon. Note that the sufficient defense capability showed the advantage compared to insufficient and none capabilities from the reliable subscale, but other subscales did not show significant differences between sufficient and insufficient. Moreover, participants highly evaluated autonomous robots in competent and ethical subscales more than the controlled robots when the robot defense capability was sufficient.

IV. Discussion

The current study investigated how people perceive the acceptability of a robot's action in self-defense scenarios based on two main factors: the level of autonomy of the robot (autonomous or remotely controlled) and the success of the robot in stopping an attacker (successful, unsuccessful, or no intervention).

- **Overall**. Our hypotheses were partially supported. We predicted that a human-controlled robot would receive higher scores in all subscales (reliable, competent, ethical, transparent, and benevolent). The results only showed a

TABLE I
MDMT SUBSCALE (AVERAGE AND S.E.) FOR EACH DEFENSE AND CONTROL CONDITION

Defense	Control	Reliable	Competent	Ethical	Transparent	Benevolent
None	Autonomous	2.736 (0.164)	2.532 (0.155)	2.977 (0.178)	3.878 (0.143)	2.765 (0.190)
	Operator	2.898 (0.153)	2.556 (0.153)	2.977 (0.182)	3.807 (0.147)	2.728 (0.180)
Insufficient	Autonomous	3.731 (0.167)	3.486 (0.158)	4.264 (0.181)	4.900 (0.145)	4.564 (0.193)
	Operator	3.952 (0.156)	3.577 (0.156)	4.226 (0.185)	4.842 (0.150)	4.365 (0.184)
Sufficient	Autonomous	4.486 (0.164)	4.292 (0.155)	4.394 (0.178)	4.933 (0.143)	4.167 (0.190)
	Operator	4.574 (0.153)	4.037 (0.153)	3.861 (0.182)	4.715 (0.147)	3.747 (0.180)

preference for the operated robot in the reliable subscale, while autonomous robots were preferred in two subscales (ethical and transparent) subscales. Furthermore, we also predicted a higher impact on people's preferences between different levels of capability. However, the results only showed a significant difference in the reliable subscale when comparing sufficient and insufficient capabilities.

- **H1 Control Factor**. We initially predicted that a remotely-controlled robot would be perceived as better performing and moral. While the participants considered the remotely-controlled robot to be more reliable, it is noteworthy that the autonomous robot was considered more competent and ethical when able to stop the aggression. Although we initially considered that, in Japan, people are more likely to trust law enforcement compared to the US, it is possible that there is a similar feeling of distrust and doubt in other humans (i.e., the operator) regardless of whether they are part of law enforcement or not (which is not clear in the videos shown). Meanwhile, an autonomous robot could be seen as someone impartial, objective, and algorithmic. Popular robotic characters in science fiction (e.g., Data from Star Trek: The Next Generation, Robocop from the RoboCop franchise, or Mira from Ghost in the Shell) might also influence participants to consider autonomous robots righteous and fair.[10,11,12] Thus, H1 is partially supported. It is also possible that the mainstream coverage of ChatGPT has affected how the participants perceive an autonomous robot, as they are more exposed and used to a complex AI application.[13]
- **H2 Capability Factor**. The results supported all three sub-hypotheses. The robot that successfully defended the victim was perceived as better performing than the robot that did not defend or the robot that tried but failed to defend, but not more moral. Furthermore, the robot that tried but failed to defend was perceived as more ethical and benevolent. Thus, H2 was fully supported.

The results seem to suggest that participants may prefer a human operator to be in control of the robot's actions for specific scenarios, but the preference for autonomy might be stronger in other scenarios where human intervention might not be feasible or effective. Even though self-defense robots are still considered science fiction, it was interesting to note that the surreality of these circumstances did not negatively affect the participants' trust in autonomous robots.

[10]en.wikipedia.org/wiki/Star_Trek:_The_Next_Generation
[11]en.wikipedia.org/wiki/RoboCop_(franchise)
[12]en.wikipedia.org/wiki/Ghost_in_the_Shell
[13]chat.openai.com

A. Limitations and Future Work

The results are limited by various factors, including the country tested and number of participants (180 participants in Japan), and the nature of the study (an online survey with animation videos). Next steps will involve gaining more insight into effects of potentially important factors such as age, gender, and culture, as well as by exploring other scenarios (e.g., determining if participants feel the same when a high degree of immersion is experienced via a virtual reality headset). As well, much technical work remains to achieve self-defense robots, both tele-operated or autonomous, in reality. By doing so, the aim is to help support a safer future in which trusted robots can help people in need.

ACKNOWLEDGMENT

We are grateful to our participants.

REFERENCES

[1] H.-P. Yueh, W. Lin, S.-C. Wang, and L.-C. Fu, "Reading with robot and human companions in library literacy activities: A comparison study," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1884–1900, 2020.

[2] L. Hung, M. Gregorio, J. Mann, C. Wallsworth, N. Horne, A. Berndt, C. Liu, E. Woldum, A. Au-Yeung, and H. Chaudhury, "Exploring the perceptions of people with dementia about the social robot paro in a hospital setting," *Dementia*, vol. 20, no. 2, pp. 485–504, 2021.

[3] K. S. Haring, M. M. Novitzky, P. Robinette, E. J. De Visser, A. Wagner, and T. Williams, "The dark side of human-robot interaction: ethical considerations and community guidelines for the field of hri," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 689–690.

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

[5] 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.

[6] ——, "Robot self-defense: Robots can use force on human attackers to defend victims," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 1606–1613.

[7] M. Cooney, M. Shiomi, E. K. Duarte, and A. Vinel, "A broad view on robot self-defense: Rapid scoping review and cultural comparison," *Robotics*, vol. 12, no. 2, p. 43, 2023.

[8] D. Gallimore, J. B. Lyons, T. Vo, S. Mahoney, and K. T. Wynne, "Trusting robocop: Gender-based effects on trust of an autonomous robot," *Frontiers in Psychology*, vol. 10, p. 482, 2019.

[9] D. Ullman and B. F. Malle, "Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 618–619.

[10] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 251–258.

[11] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[12] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 2016, pp. 101–108.

[13] A. Weiss, D. Wurhofer, M. Lankes, and M. Tscheligi, "Autonomous vs. tele-operated: How people perceive human-robot collaboration with hrp-2," in *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2009, pp. 257–258.

[14] J. Nasir, P. Oppliger, B. Bruno, and P. Dillenbourg, "Questioning wizard of Oz: Effects of revealing the wizard behind the robot," in *2022 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, no. CONF, 2022.

[15] D. Tozadore, A. Pinto, R. Romero, and G. Trovato, "Wizard of Oz vs autonomous: Children's perception changes according to robot's operation condition," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 664–669.

[16] M. Bennett, T. Williams, D. Thames, and M. Scheutz, "Differences in interaction patterns and perception for teleoperated and autonomous humanoid robots," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 6589–6594.

[17] M. A. Goodrich, A. C. Schultz *et al.*, "Human–robot interaction: a survey," *Foundations and Trends® in Human–Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2008.

[18] A. Sharkey and N. Sharkey, "Granny and the robots: ethical issues in robot care for the elderly," *Ethics and information technology*, vol. 14, pp. 27–40, 2012.

[19] K. Zheng, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "How many social robots can one operator control?" in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2011, pp. 379–386.

[20] 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.

[21] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[22] A. M. Aroyo, F. Rea, G. Sandini, and A. Sciutti, "Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?" *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3701–3708, 2018.

[23] M. K. Lee, S. Kiesler, and J. Forlizzi, "Receptionist or information kiosk: how do people talk with a robot?" in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 2010, pp. 31–40.

[24] S. You and L. P. Robert Jr, "Human-robot similarity and willingness to work with a robotic co-worker," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 251–260.

[25] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018.

[26] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 141–148.