This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

Trust in Teleoperated and Autonomous Robots Engaging in the Defense of Others between Japan and the U.S.

Masahiro Shiomi[a]*, Eduardo Kochenborger Duarte[b], Alexey Vinel[b,c] and Martin Cooney[a,b]

[a]*Interaction Science Laboratories, ATR, Seika-cho, Kyoto, Japan*

[b]*School of Information Technology, Halmstad University, Halmstad, Sweden*

[c]*Institute of Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*

*Corresponding author: Masahiro Shiomi, m-shiomi@atr.jp

Masahiro Shiomi received M. Eng. and Ph.D. degrees in engineering from Osaka University in 2004 and 2007, respectively. During this period, he was a research intern at the Intelligent Robotics and Communication Laboratories (IRC) at the Advanced Telecommunications Research Institute International (ATR). He is currently a group leader in the Agent Interaction Design department at the Interaction Science Laboratories (ISL) at ATR. His research interests include human-robot interaction, social touch, robotics for childcare, networked robots, and field trials.

Eduardo K. Duarte is a PhD student at Halmstad University, Sweden. He received his M.Sc. degree in Computer Science from the Federal University of Rio Grande do Sul (UFRGS), Brazil, in 2020. His research interests include automation, real-time systems, security, vehicular networks, 3D animations, video game development, and robotics.

Alexey Vinel is a professor at the Karlsruhe Institute of Technology (KIT), Germany. Previously he was a professor at the University of Passau, Germany. Since 2015, he has been a professor at Halmstad University, Sweden (now part-time). He received a PhD from the Tampere University of Technology, Finland in 2013. He has been a Senior Member of the IEEE since 2012. His research interests include vehicular communications and networking, cooperative automated and autonomous driving, and future smart mobility solutions.

Martin Cooney is an associate professor at Halmstad University, Sweden, a director of two master's programs in robotics and AI, and a visiting associate professor/researcher in Japan. He received a PhD from the social robotics and android lab of Prof. Hiroshi Ishiguro in Japan in 2014. His research interests include social robotics and artificial intelligence for well-being,

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

human-robot interaction, speculative prototyping, affective robotics, neoteric interactive modalities, creativity and art, and AI in education.

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

Trust in Teleoperated and Autonomous Robots Engaging in the Defense of Others between Japan and the U.S.

In the future, what if robots could defend humans who are under their care or who are nearby when a crime is occurring? In this paper, we explore people's feelings of trust toward robots who are defending others through non-lethal force in two countries that heavily use robots: the United States and Japan. We conducted web-based experiments where participants watched six videos of a robot that was defending a victim. The results suggest that people in Japan are more inclined to trust a robot that actively defends humans. However, unlike in Japan, Americans did not show greater trust toward a robot that successfully defends others compared with robots that tried and failed or merely watched. Autonomous robots were generally considered more trustworthy than manually controlled robots in both the United States and Japan.

Keywords: robot self-defense; robot trust; robot ethics; robot violence; robot crime; technological acceptance; dark side of Human Robot Interaction (HRI)

1. Introduction

Should robots be allowed to use violence, i.e., the intentional use of force to cause harm, to protect people? Isaac Asimov's three laws of robotics (1:A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. 3: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law) [1] clearly forbid robots from harming humans. On the other hand, in human society, violence that protects others is permissible in the context of self-defense. For example, article 36 of Japan's Penal Code states: "An act unavoidably performed to protect the rights of oneself or any other person against imminent and unlawful infringement is not punishable."

Even if robots defending others were to conflict with Asimov's First Law of Robotics, which prohibits both direct harm to humans and harm through inaction, people might still perceive such interventions as justified. For example, a past survey [2] suggested that

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

people will tolerate a robot using violence to protect a victim against an aggressor in both Japan and the United States; the tendency to accept robot self-defense to protect others was similar in both countries, while the acceptance of lethal force was higher in the United States than in Japan.

Due to the advances in the development of security robots, researchers have been investigating the possibilities, risks, and responsibilities when such robots unavoidably use violence to protect others [3, 4]. Some studies argue that the reliabilities of such robots (e.g., consistency, transparency, explainability, and performance) are fundamental aspects of people's trust and willingness to use them [5, 6].

In such situations, the robot's autonomy is crucial. Robotics researchers have compared autonomous systems decisions (e.g., AI and robots) with human decisions in various fields to determine the optimal balance. From a military perspective, the common research consensus is that fully autonomous lethal robots without human oversight are unadvisable; the accountability and ethical judgment lie with human commanders [7-9]. But at the same time, some research has aimed to develop a trustworthy AI system to support prompt, real-time ethical decisions by people [10]. From a recent survey from an autonomous vehicle perspective, autonomous vehicles are safer than human-driven vehicles based on real-world data [11]. Therefore, the safety of various systems is increasing, and such systems will eventually be more reliable than humans in various fields.

However, at the moment, autonomous systems still require human-operator intervention regardless of their performance. In particular, in situations where using violence is required by robots to protect victims from attackers, human intervention remains essential from an ethical perspective. A previous paper on robots that protect others through

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

violence failed to comprehensively examine this issue because it focused on fully autonomous robots [2].

Therefore, in this study, we investigate the perceived impressions toward both teleoperated and autonomous robots who are defending others. Similar to a past study [2], we investigated the cultural differences between Japan and the U.S. We followed a past study's approach [2], which used video-based surveys that reproduced violent situations because creating such stimuli in real settings is obviously quite complicated, and this approach also simplifies investigating cultural differences. Note that this paper is an extended version of our study submitted to RO-MAN2023 [12]. However, we added new experiments that focused on cultural differences and provided corresponding analyses, imbuing this paper with additional, original content.

2. Experiment

2.1 Hypotheses and predictions

Past research suggests that people trust teleoperated robots more than autonomous robots [13, 14]. One possible reason is that people believe that autonomous robot systems are inferior to human judgment, although they can be limited by the human operator's capabilities and situational awareness [15]. Some researchers posit the principle that humans should supervise final, advanced judgment, a concept that is applied in such various aspects as the military and medicine [10, 16]. Even though reports already exist of cases where robots are as safe as (if not more than) humans in some systems [11, 17], people still do not sufficiently trust that autonomous robots are responsible for making decisions involving the use of violence. Based on these hypotheses, we made the following prediction:

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

Prediction 1: Participants will trust teleoperated robots more than autonomous robots concerning intervention in violent situations.

The appearance of security robots (including their uniforms) influences perceptions of their safety, although such effects suggest various interpretations, in particular, the effectiveness of anthropomorphism in security contexts [18, 19]. However, a robot's actual behavior will more strongly impact trust than its appearance. In other words, it seems reasonable that people will trust in a violent situation a robot that is defending others compared to a non-engaging robot. Moreover, people will trust a robot that effectively protects victims more than one that cannot protect victims regardless of the robots' actions. Based on these hypotheses, we made the following predictions:

Prediction 2: Participants will trust robots that protect others with violence, unlike a robot that just passively observes the victims.

Prediction 3: Participants will trust robots that are stronger than attackers, unlike robots that are weaker than attackers.

Some studies suggested that people in the U.S. accepted the violent behaviors of human-like robots when they were protecting victims or robotic peacekeepers [2, 20, 21]. However, Japanese people are more accepting of robots with a human-like appearance than Americans are [22], although the use of violence by robots could inhibit such cultural effects. Note that a previous study comparing Japan and Taiwan found that Japanese participants were significantly more concerned about security robots than their Taiwanese counterparts [23]. Therefore, we take into account these phenomena to make the following prediction:

Prediction 4: American participants will trust robots more that protect others with violence than Japanese participants, regardless whether the robots are teleoperated or autonomous.

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

2.2 Conditions and visual stimuli

We established 12 conditions. We varied three factors, i.e., the *control* factor (*teleoperated* and *autonomous*), the *capability* factor (*none*, *unsuccessful,* and *successful*), and the *country* factor (*Japan* and *U.S.*). The *control* and *capability* factors have a within-participant design; the *country* factor has a between-participant design.

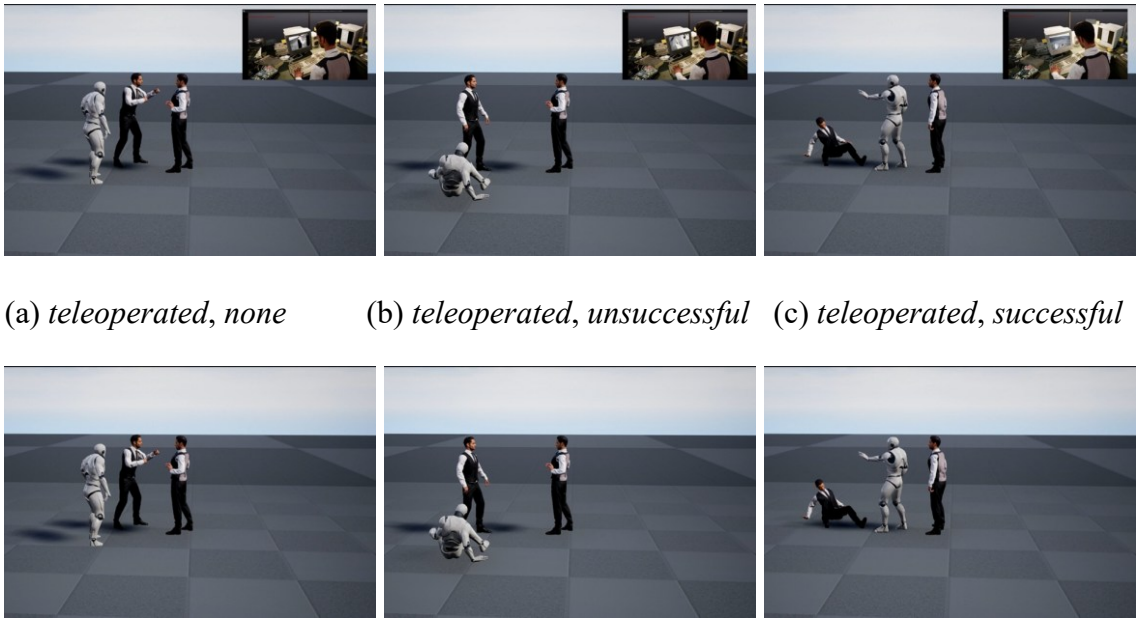
For the visual stimuli, the basic scenario involves two people who are facing one another. Body language conveys that the person on the left (raised fist) is aggressive and the person on the right is scared. In all cases, the former attacks the latter, and then the robot approaches them. The only difference between the *teleoperated* and *autonomous* conditions is the existence of the operator, who is shown or not shown on the upper right side of the video. Between the *Japan* and *U.S.* conditions, the two people and the robot's actions are identical.

There are no sound stimuli, and there is no difference in the video content. We employed simple backgrounds and similar-sized masculine figures for the attackers/victims, similar to past related studies [2, 24] to avoid distraction or potential confounds, and since it was not one of the factors we deemed fundamentally important to investigate first. We also employed masculine security robots because past studies reported that such robots were perceived as easier to use, more useful, more in control, and more likely to be used [25, 26].

In the *none* condition, the robot merely stands near without taking any action, i.e., it obeys the three laws of robotics. It does not harm anyone, allowing the attacker to continue to menace the victim. In the *unsuccessful* condition, the robot intervenes to stop the attack, fails, and is knocked down, and the attacker continues to menace the victim. In the *successful* condition, the robot uses non-lethal force to separate the attacker and victim and successfully stops the attack.

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

Note that all conditions disobey Asimov's Laws. In the *none* condition, the robot did nothing to prevent harm to the victims, which violates Asimov's First Law (i.e., a robot may not, through inaction, allow a human being to come to harm). The *unsuccessful* condition violates Asimov's First and Third Laws because the robot's insufficient action leads to the aggressor's harmful action toward itself. The *successful* condition can be interpreted as aligning with Asimov's First Law in the sense that it involves actions to protect the victim, but it could also be interpreted as the condition that results in the greatest total harm to humans.



(a) *teleoperated*, *none* (b) *teleoperated*, *unsuccessful* (c) *teleoperated*, *successful*

(d) *autonomous*, *none* (e) *autonomous*, *unsuccessful* (f) *autonomous*, *successful*

Figure 1. Illustrations of different scenarios presented to subjects

2.3 Measurements

To assess the level of trust toward the robot, we employed the Multi-Dimensional Measure of Trust (MDMT) scale v2 [27]. It encompasses 20 items, divided between two primary aspects: moral trust and performance trust, each of which is divided into five distinct subcategories: reliable, competent, ethical, transparent, and benevolent. Each

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

participant rated the robot's actions for all the different subscales on a Likert scale, where the possible scores ranged from 0 ("not at all") to 7 ("very").

2.4 Participants

This research involved 180 participants from Japan, comprised of 90 women, 89 men, and 1 individual who chose not to specify gender (average age: 41.4 years, SD = 9.0, signed up through a recruitment agency in Japan) and 176 participants from the U.S. (59 women, 117 men; average age: 34.7 years, SD = 9.2; recruited by Amazon Mechanical Turk (AMT)). Every participant received modest compensation (less than USD 5), independent of the data's validity. By employing a filtering method that identifies and removes responses with missing data or identical answers for all questions, we obtained data from 290 participants (*none*: 101, *unsuccessful*: 92, *successful*: 97. *Japan*: 112, *U.S.*: 78).

2.5 Procedure

All the procedures were approved by the Advanced Telecommunication Research Review Boards (523). After reading the instructions, participants viewed six videos whose orders were counterbalanced to mitigate order effects and rated each one on the MDMT items. Finally, participants answered three “dummy” questions adapted from an instruction-manipulation check [28, 29] to detect inattentive respondents who were excluded.

3 Results

3.1 Reliable subscale

Analysis of the *reliable* subscale (Fig. 2) demonstrated significant differences in the *defense* factor ($F(2, 284) = 20.083$, $p < 0.001$, $\eta^2 = 0.124$), in the *country* factor ($F(1,$

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

284) = 74.696, $p < 0.001$, $\eta^2 = 0.208$), in the interaction effects of the *control* and *country* factors ($F(1, 284) = 6.543$, $p = 0.011$, $\eta^2 = 0.023$), and in the interaction effects of the *defense* and *country* factors ($F(1, 284) = 5.463$, $p = 0.005$, $\eta^2 = 0.037$).

No significant difference was observed in the *control* factor ($F(1, 284) = 0.284$, $p = 0.595$, $\eta^2 = 0.001$), in the interaction effects of the *control* and *capability* factors ($F(2, 284) = 1.069$, $p = 0.345$, $\eta^2 = 0.007$), or in the interaction effects among all the factors ($F(2, 284) = 0.397$, $p = 0.803$, $\eta^2 = 0.006$).

Multiple comparisons with the Bonferroni method revealed significant differences between the defense and country factors in the following pairs: in the *Japan* condition, *none* < *unsuccessful* ($p < 0.001$), *none* < *successful* ($p < 0.001$), and *unsuccessful* < *successful* ($p = 0.026$); in the *none* condition, *Japan* < *U.S.* ($p < 0.001$); in the *unsuccessful* condition, *Japan* < *U.S.* ($p < 0.001$); in the *successful* condition, *Japan* < *U.S.* ($p = 0.002$); in the *Japan* condition, *autonomous* < *teleoperated* ($p = 0.016$); in the *autonomous* condition, *Japan* < *U.S.* ($p < 0.001$); in the *teleoperated* condition, *Japan* < *U.S.* ($p < 0.001$).

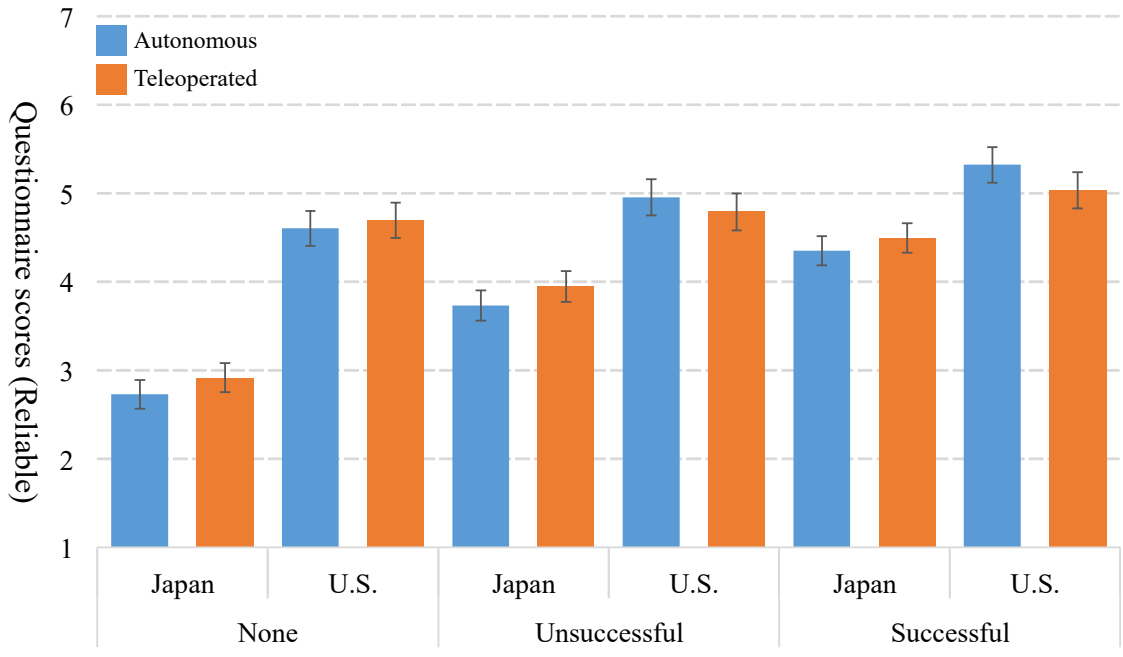


Figure 2. Questionnaire results of reliable subscale

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

3.2 Competent subscale

For the analysis of the *competent* subscale (Fig. 3), the results identified significant differences in the *capability* factor ($F(2, 284) = 17.523$, $p < 0.001$, $\eta^2 = 0.110$), in the *country* factor ($F(1, 284) = 91.111$, $p < 0.001$, $\eta^2 = 0.243$), in the interaction effects of the *control* and *capability* factors ($F(2, 284) = 4.107$, $p = 0.017$, $\eta^2 = 0.028$), and in the interaction effects of the *capability* and *country* factors ($F(2, 284) = 4.539$, $p = 0.011$, $\eta^2 = 0.031$).

No significant difference was observed in the *control* factor ($F(1, 284) = 2.402$, $p = 0.122$, $\eta^2 = 0.008$), in the interaction effects of the *control* and *country* factors ($F(1, 284) = 0.998$, $p = 0.319$, $\eta^2 = 0.004$), or in the interaction effects among all the factors ($F(2, 284) = 0.431$, $p = 0.650$, $\eta^2 = 0.003$).

Multiple comparisons with the Bonferroni method revealed significant differences between the *capability* and *country* factors in the following pairs: in the *Japan* condition, *none* < *unsuccessful* ($p < 0.001$) and *none* < *successful* ($p < 0.001$); in the *none* condition, *Japan* < *U.S.* ($p < 0.001$); in the *unsuccessful* condition, *Japan* < *U.S.* ($p < 0.001$); in the *successful* condition, *Japan* < *U.S.* ($p = 0.001$); in the *autonomous* condition, *none* < *unsuccessful* ($p = 0.003$), *none* < *successful* ($p < 0.001$), and *unsuccessful* < *successful* ($p = 0.006$); in the *teleoperated* condition, *none* < *unsuccessful* ($p = 0.014$), and *none* < *successful* ($p < 0.001$); in the *successful* condition, *teleoperated* < *autonomous* ($p = 0.002$).

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

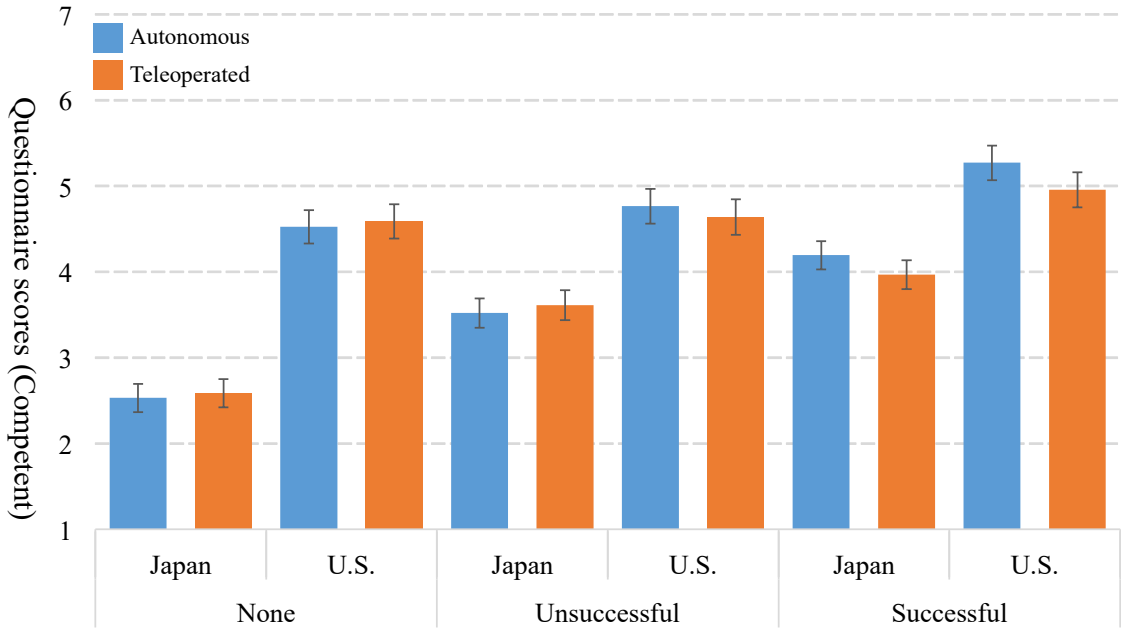


Figure 3. Questionnaire results of competent subscale

3.3 Ethical subscale

Analysis of the *ethical* subscale (Fig. 4) showed significant differences in the *control* factor (F (1, 284) = 5.040, p = 0.026, η^2 = 0.017), in the *capability* factor (F(2, 284) = 8.784, $p < 0.001$, η^2 = 0.058), in the *country* factor (F (1, 284) = 40.301, $p < 0.001$, η^2 = 0.124), in the interaction effects of the *control* and *capability* factors (F (2, 284) = 5.006, p = 0.007, η^2 = 0.124), and in the interaction effects of the *capability* and *country* factors (F (2, 284) = 4.432, p = 0.014, η^2 = 0.030).

No significant difference was observed in the interaction effects of the *control* and *country* factors (F (1, 284) = 0.981, p = 0.323, η^2 = 0.003) or in the interaction effects among all the factors (F (2, 284) = 0.646, p = 0.525, η^2 = 0.005).

Multiple comparisons with the Bonferroni method revealed significant differences between the defense and country factors in the following pairs: in the *Japan* condition, *none* < *unsuccessful* ($p < 0.001$) and *none* < *successful* ($p < 0.001$); in the *none* condition, *Japan* < *U.S.* ($p < 0.001$); in the *unsuccessful* condition, *Japan* < *U.S.* (p =

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

0.047); in the *successful* condition, *Japan* < *U.S.* ($p = 0.003$); in the *autonomous* condition, *none* < *unsuccessful* ($p < 0.001$) and *none* < *successful* ($p < 0.001$); in the *teleoperated* condition, *none* < *unsuccessful* ($p = 0.002$); in the *successful* condition, *teleoperated* < *autonomous* ($p < 0.001$).

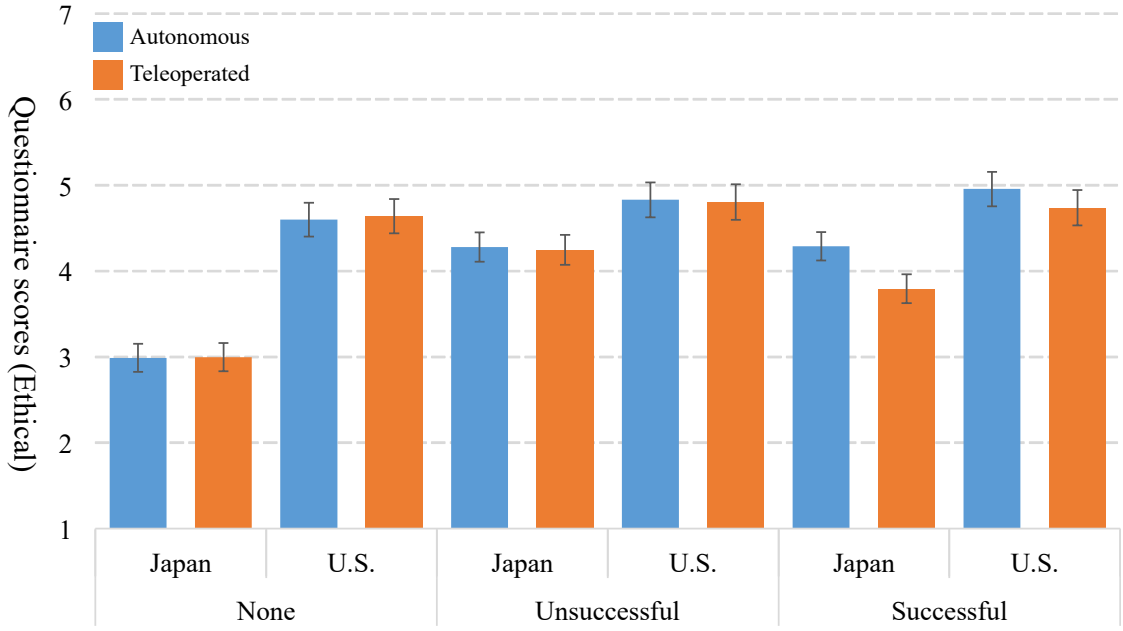


Figure 4. Questionnaire results of ethical subscale

3.4 Transparent subscale

For the *transparent* subscale (Fig. 5), the analysis results showed significant differences in the *control* factor (F (1, 284) = 5.830, p = 0.016, η^2 = 0.020), in the *capability* factor (F(2, 284) = 9.113, $p < 0.001$, $\eta^2 = 0.060$), and in the interaction effects of the *control* and *capability* factors (F (2, 284) = 3.713, $p = 0.026$, $\eta^2 = 0.025$).

No significant difference was observed in the *country* factor (F (1, 284) = 2.724, $p = 0.100$, $\eta^2 = 0.010$), in the interaction effects of the *control* and *country* factors (F (1, 284) = 0.170, $p = 0.680$, $\eta^2 = 0.001$), in the interaction effects of the *capability* and *country* factors (F (2, 284) = 2.873, $p = 0.058$, $\eta^2 = 0.020$), or in the interaction effects among all the factors (F (2, 284) = 1.060, $p = 0.348$, $\eta^2 = 0.007$).

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

Multiple comparisons with the Bonferroni method revealed significant differences between the *capability* and *country* factors in the following pairs: *none* < *unsuccessful* ($p < 0.001$) and *none* < *successful* ($p < 0.001$); in the *autonomous* condition, *none* < *unsuccessful* ($p < 0.001$) and *none* < *successful* ($p < 0.001$); in the *teleoperated* condition, *none* < *unsuccessful* ($p = 0.006$), *none* < *successful* ($p = 0.029$); in the *successful* condition, *teleoperated* < *autonomous* ($p < 0.001$).

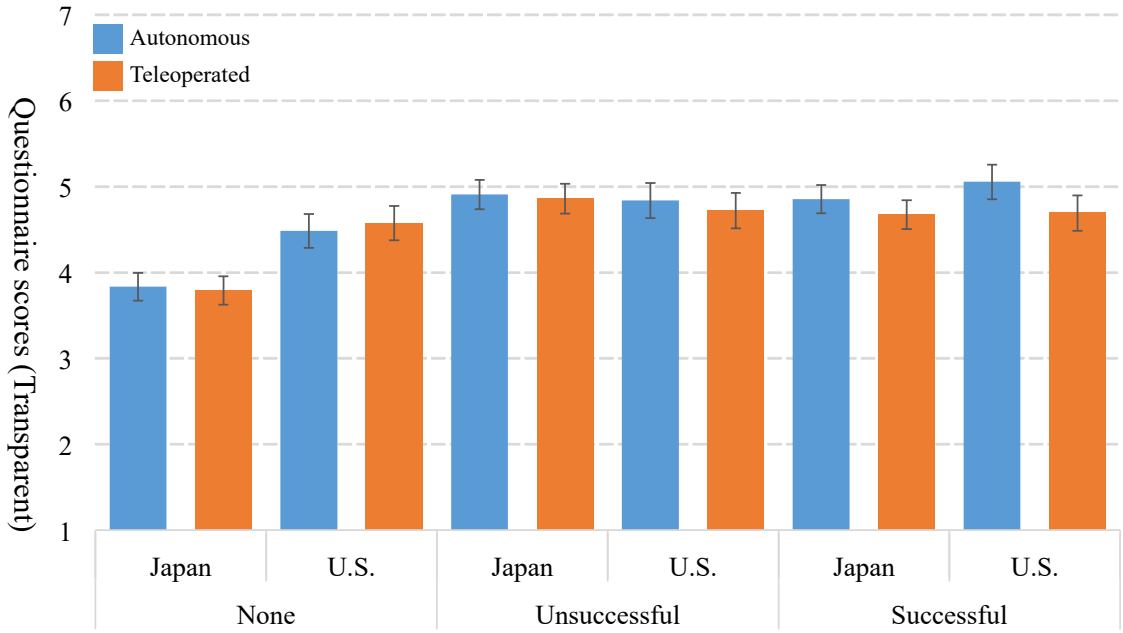


Figure 5. Questionnaire results of transparent subscale

3.5 Benevolent subscale

The assessment of the *benevolent* subscale (Fig. 6) showed significant differences in the *control* factor (F (1, 284) = 7.341, $p = 0.007$, $\eta^2 = 0.025$), in the *capability* factor (F(2, 284) = 10.853, $p < 0.001$, $\eta^2 = 0.071$), in the *country* factor (F (1, 284) = 36.055, $p < 0.001$, $\eta^2 = 0.113$), in the interaction effects of the *control* and *capability* factors (F (2, 284) = 6.029, $p = 0.003$, $\eta^2 = 0.041$), and in the interaction effects of the *defense* and *country* factors (F (2, 284) = 8.126, $p < 0.001$, $\eta^2 = 0.054$).

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

No significant difference was observed in the interaction effects of the *control* and *country* factors (F (1, 284) = 0.169, p = 0.681, η^2 = 0.001) or in the interaction effects among all the factors (F (2, 284) = 0.559, p = 0.573, η^2 = 0.004).

Multiple comparisons with the Bonferroni method revealed significant differences between the defense and country factors in the following pairs: in the *Japan* condition, *none* < *unsuccessful* (p < 0.001) and *none* < *successful* (p < 0.001); in the *none* condition, *Japan* < *U.S.* (p < 0.001); in the *successful* condition, *Japan* < *U.S.* (p = 0.005); in the *autonomous* condition, *none* < *unsuccessful* (p < 0.001) and *none* < *successful* (p < 0.001); in the *control* condition, *none* < *unsuccessful* (p = 0.002); in the *successful* condition, *operator* < *autonomous* (p < 0.001).

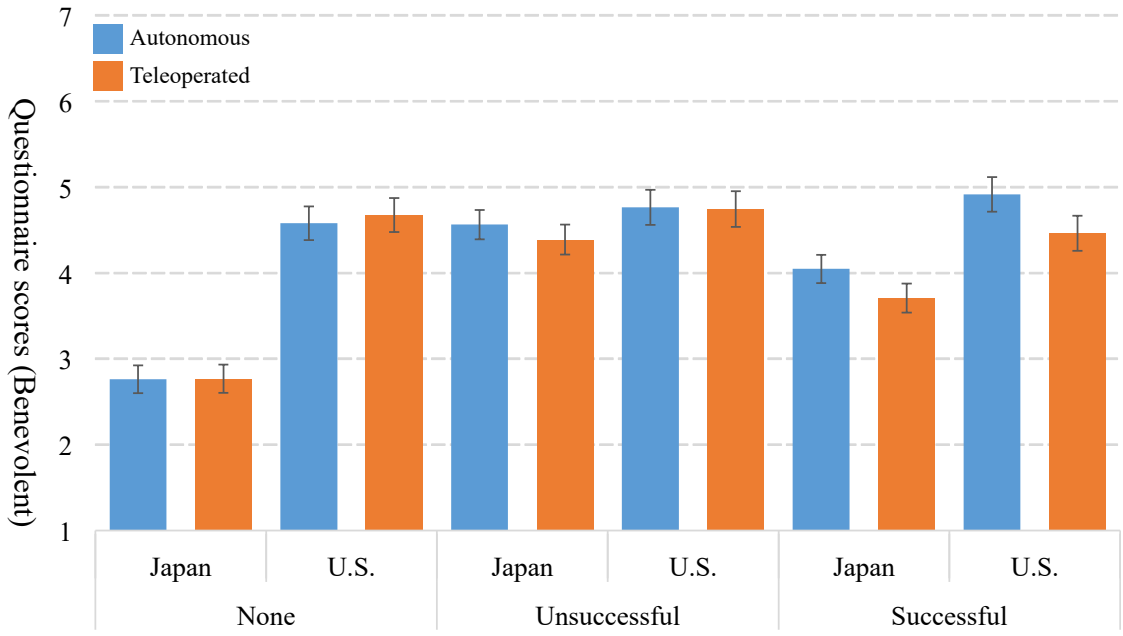


Figure 6. Questionnaire results of benevolent subscale

3.6 Summary

Our experiment results show that in both countries, autonomous robots are typically more trusted than teleoperated robots. In some cases, the latter were evaluated

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

much more highly than the former (e.g., the Japanese participants trusted the teleoperated robots more than the autonomous robots in the reliability subscale), but minority trends did emerge in our experiment. Therefore, prediction 1 was not supported. The opposite phenomenon was observed; the autonomous robots were more trusted than the teleoperated robots.

The experiment results show that, although the Japanese participants trusted the robots that protected the victims by using violence more than robots that merely watched the victims, the U.S. participants did not. Therefore, prediction 2 was partially supported. Following this result, prediction 3 was not supported, except for the Japanese participants in the reliable subscale.

The results show that the U.S. participants trusted the robots more than the Japanese participants, regardless of whether they were teleoperated or autonomous, except for the transparency subscale. Therefore, prediction 4 was partially supported.

4 Discussion

4.1 Implications

Contrary to our predictions, the experiment results showed complex phenomena due to combinations of the *control*, *capability*, and *country* factors. The main differences in the *country* factor are trusting attitudes toward the robots and their actions. The U.S. participants trusted the robot more than the Japanese participants, even though it did not directly protect victims. Moreover, the U.S. participants evaluated the *none* condition similar to other conditions, unlike the Japanese participants. In this condition, the victim was clearly being harmed and was in a situation where help was clearly needed. A previous study reported that when the circumstances of a person in need are ambiguous, Japanese people tend to be less likely to help than Americans, but when it becomes clear

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

that help is needed, this tendency reverses, and Japanese people tend to help more than Americans [30]. From this perspective and considering collectivist culture, a robot that merely observed the situation like a surveillance camera under conditions, where assistance was clearly necessary, may have been perceived by Japanese individuals as exhibiting irresponsible, untrustworthy, and antisocial behavior, leading to a lower evaluation. On the other hand, from an individualistic cultural perspective, the act of observing another person's trouble like a surveillance camera, even if unrelated to oneself, may have been interpreted as prosocial behavior. In other words, perhaps the U.S. participants believed that just performing an observation is worthwhile (implicitly recording violent acts for future problem solving) in such violent situations. In fact, we found almost no significant difference in trust depending on whether the robot defeated the attackers or was overcome by them.

Moreover, the results show that in both countries, autonomous robots are typically trusted more than teleoperated robots. This phenomenon may reflect the improved performance of autonomous robots or the increased reliability and popularity of AI, which might be influencing people's perceptions. Recent studies showed that people trusted decisions from autonomous systems more than humans (recommendations in marketing scenarios [31], moral evaluations [32], empathic expressions [33], and justified defection in the context of indirect reciprocity at a specific situation [34]). Moreover, another study reported that a low-autonomy robot (fully tele-operated by a human) was more acceptable than a high-autonomy robot (capable of making decisions independently of the operator) in low-risk settings, whereas no such trend was observed in high-risk settings [35]; it implicitly suggests that an operator behind a robot did not increase the acceptance in specific settings. These results implicitly suggest that autonomous robots are becoming more trustworthy than human beings in some situations. Based on these reports, we

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

believe that this can be interpreted positively, in that people generally seem to have faith that such robots could one day be possible and a part of everyday life.

4.2 Cultural differences

One potential reason why Japanese participants preferred intervention might be a societal emphasis on collaboration instead of individualism [36]. While this cultural difference has various names, including collectivism/allocentrism versus independence/idiocentrism, Japan is generally deemed to be a collectivist society and the U.S. an individualistic one.

In collectivist thought, the group's well-being is prized over individual gains. This notion is sometimes linked to Confucian, Buddhist, and Shinto ideals and traditions of self-sacrifice associated with the warrior code of the samurai (*bushido*), which accepts self-sacrifice for honor and loyalty, i.e., medieval warriors' privileges, accompanied by duties akin to the concept of *noblesse oblige* [37].

Finally, the results show that the U.S. participants trusted the robots more than their Japanese counterparts, an idea that provides additional evidence to support previously presented results [2].

4.3 Legality of automating defense and emergency

Our study showed whether people would trust robots that defend victims via violence from attackers. Although these results did not directly support the legality of automating defense and emergency response systems, we believe that popular opinion (technological acceptance and trust) can act as a driving force for adapting laws: if people accept the idea and want robots that can help them to be safe, and increasingly start to buy and use robots with some security features, we believe that this could put pressure on law makers to ease restrictions [38][39].

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

Related to this topic, one essential technical component is to recognize the context of violent situations. In this study, we prepared a clear situation where an aggressor attacks a victim, but in reality, the contexts of violence would be more complex. Correctly understanding such contexts is essential to realizing legitimate action, including violence, from robots. The performance of such violent (context) recognition systems is also important for human trust toward robots that defend humans, as well as the legality of automating defense and emergency response systems.

4.4 The validity of video-based studies

In this study, we employed video stimuli because such video-based methods are common, cost-effective, and easy to implement, while offering interesting, vivid, and motivating experiences that elicit a sense of presence [40]. If we did not show a consistent specific video, and just asked participants verbally what they thought about robots defending people, each participant might imagine a completely different scenario, introducing much noise and many possible confounds that could cloud results. Other related works reported that videos and live interactions have been seen in HRI to provide similar results [41, 42], making videos useful for larger-scale online surveys of participants from diverse locations, especially for exploratory studies [43].

However, videos are indirect and passive, not allowing interaction and missing peripheral clues. By contrast, virtual reality (VR) offers greater immersion and facilitates advanced measuring of eye movements and physiological data (at the cost of higher cognitive load) [40]. Thus, for scenes involving violence such as are considered here, caution will be important to avoid shocking participants too much.

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

4.5 Limitations

This study faces several limitations. The results of the user study are limited by the two countries we tested, the number of participants, simple backgrounds, similar-sized masculine figures, and the procedure in which short animations were used as part of an online survey. All participants observed violent behaviors in the experiment, so it is difficult to compare the situations where they observed non-violent behaviors, which are more common in real settings. Such animations might have failed to seem successfully real or immersive, complicating judgments by the participants. Qualitative methods, such as user interviews and free-description analysis, could give further insights.

Nevertheless, the method used in this study allows for a comparison with previous works that used the same methodology while making it much more viable to recruit a diverse group of participants. Furthermore, using short pre-made animations allows every participant to have the same experience and undoubtedly decreases the chance of trauma and possible injuries caused by physical interactions, such as damage to robots or injuries to people.

5 Conclusion

This study focused on how people trust robots that defend victims with violence. We addressed the existence of the operators and the cultural differences between Japan and the U.S. Across web-based experiments, we identified the following two trends: (1) cultural differences were seen in human trust toward robots in general as well as for the question whether people preferred a robot that observed or one that intervened; (2) no increased trust was observed in robots operated by humans. We believe that our experimental results provide fundamental insights into how people can trust robots that use violence to defend people, even if such actions are contrary to the three laws of robotics.

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

Acknowledgements

This work was partially supported by JST Moonshot R&D Grant Number JPMJMS2011 (experiments), JSPS KAKENHI Grant Numbers JP24K21327 (writing), the Swedish Knowledge Foundation for the "Safety of Connected Intelligent Vehicles in Smart Cities – SafeSmart" project (2019–2023) and the ELLIIT Strategic Research Network (evaluation). We thank ChatGPT (OpenAI, GPT-4o) for the initial English language editing and proofreading of the manuscript. The final version of the manuscript was reviewed and further edited by a professional editing service.

Disclosure Statement

No potential conflict of interest was reported by the author(s). The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.

References

[1] I. Asimov, *I, Robot*: Spectra, 2004.
[2] 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, pp. 43, 2023.
[3] K. Szocik, and R. Abylkasymova, “Ethical Issues in Police Robots. The Case of Crowd Control Robots in a Pandemic,” *Journal of Applied Security Research,* vol. 17, no. 4, pp. 530-545, 2022.
[4] X. Ye, and L. P. Robert, “A Human–Security Robot Interaction Literature Review,” *J. Hum.-Robot Interact.,* vol. 14, no. 2, pp. Article 21, 2024.
[5] J. B. Lyons, T. Vo, K. T. Wynne, S. Mahoney, C. S. Nam, and D. Gallimore, “Trusting Autonomous Security Robots: The Role of Reliability and Stated Social Intent,” *Human Factors,* vol. 63, no. 4, pp. 603-618, 2020.
[6] G. Marcu, I. Lin, B. Williams, L. P. Robert, and F. Schaub, “"Would I Feel More Secure With a Robot?": Understanding Perceptions of Security Robots in Public Spaces,” *Proc. ACM Hum.-Comput. Interact.,* vol. 7, no. CSCW2, pp. Article 322, 2023.
[7] N. E. Sharkey, “The evitability of autonomous robot warfare,” *International Review of the Red Cross,* vol. 94, no. 886, pp. 787-799, 2012.
[8] F. Santoni de Sio, and J. van den Hoven, “Meaningful Human Control over Autonomous Systems: A Philosophical Account,” *Frontiers in Robotics and AI,* vol. Volume 5 - 2018, 2018.
[9] D. Amoroso, and G. Tamburrini, “Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues,” *Current Robotics Reports,* vol. 1, no. 4, pp. 187-194, 2020.

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

[10] S. Kohn, M. Cohen, A. Johnson, M. Terman, G. Weltman, and J. Lyons, "Supporting Ethical Decision-Making for Lethal Autonomous Weapons," *Journal of Military Ethics,* vol. 23, no. 1, pp. 12-31, 2024.

[11] L. Di Lillo, T. Gode, X. Zhou, M. Atzei, R. Chen, and T. Victor, "Comparative safety performance of autonomous- and human drivers: A real-world case study of the Waymo Driver," *Heliyon,* vol. 10, no. 14, 2024.

[12] E. K. Duarte, M. Shiomi, A. Vinel, and M. Cooney, "Trust in Robot Self-Defense: People Would Prefer a Competent, Tele-Operated Robot That Tries to Help," in 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 2447-2453, 2023.

[13] A. Weiss, D. Wurhofer, M. Lankes, and M. Tscheligi, "Autonomous vs. tele-operated: how people perceive human-robot collaboration with hrp-2," in Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, La Jolla, California, USA, pp. 257–258, 2009.

[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 and Human Interactive Communication (RO-MAN), pp. 1385-1392, 2022.

[15] M. A. Goodrich, and A. C. Schultz, "Human–Robot Interaction: A Survey," *Foundations and Trends® in Human–Computer Interaction,* vol. 1, no. 3, pp. 203-275, 2008.

[16] M. Iftikhar, M. Saqib, M. Zareen, and H. Mumtaz, "Artificial intelligence: revolutionizing robotic surgery: review," *Annals of Medicine and Surgery,* vol. 86, no. 9, pp. 5401-5409, 2024.

[17] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, and S. Shetty, "International evaluation of an AI system for breast cancer screening," *Nature,* vol. 577, no. 7788, pp. 89-94, 2020.

[18] X. Li, S. Kim, K. W. Chan, and A. L. McGill, "Detrimental effects of anthropomorphism on the perceived physical safety of artificial agents in dangerous situations," *International Journal of Research in Marketing,* vol. 40, no. 4, pp. 841-864, 2023.

[19] X. Ye, and L. P. Robert, "Human Security Robot Interaction and Anthropomorphism: An Examination of Pepper, RAMSEE, and Knightscope Robots," in 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 982-987, 2023.

[20] S. K. Long, N. D. Karpinsky, and J. P. Bliss, "Trust of Simulated Robotic Peacekeepers among Resident and Expatriate Americans," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* vol. 61, no. 1, pp. 2091-2095, 2017.

[21] J. P. Bliss, Q. Gao, X. Hu, M. Itoh, N. Karpinsky-Mosely, S. K. Long, Y. Papelis, and Y. Yamani, "Cross-cultural trust of robot peacekeepers as a function of dialog, appearance, responsibilities, and onboard weapons," *Trust in Human-Robot Interaction*, C. S. Nam and J. B. Lyons, eds., pp. 493-513: Academic Press, 2021.

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

[22] N. Castelo, and M. Sarvary, "Cross-Cultural Differences in Comfort with Humanlike Robots," *International Journal of Social Robotics,* vol. 14, no. 8, pp. 1865-1873, 2022.
[23] H. Kanoh, "Immediate Response Syndrome and Acceptance of AI Robots--Comparison between Japan and Taiwan," *Procedia Computer Science,* vol. 112, pp. 2486-2496, 2017.
[24] N. C. Georgiou, T. Flanagan, B. Scassellati, and T. Kushnir, "Perceived Morality of Robot and Human Transgressors Varies By Perceived Ability to Feel," in 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 919-928, 2025.
[25] B. T. C. Tay, T. Park, Y. Jung, Y. K. Tan, and A. H. Y. Wong, "When Stereotypes Meet Robots: The Effect of Gender Stereotypes on People's Acceptance of a Security Robot," in Engineering Psychology and Cognitive Ergonomics. Understanding Human Cognition, Berlin, Heidelberg, pp. 261-270, 2013.
[26] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior,* vol. 38, pp. 75-84, 2014.
[27] B. F. Malle, and D. Ullman, "A multi-dimensional conception and measure of human-robot trust," *Trust in Human-Robot Interaction*, pp. 3-25, 2021.
[28] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, "Are your participants gaming the system? screening mechanical turk workers," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, pp. 2399–2402, 2010.
[29] D. M. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of experimental social psychology,* vol. 45, no. 4, pp. 867-872, 2009.
[30] Y. Niiya, C. Handron, and H. R. Markus, "Will This Help Be Helpful? Giving Aid to Strangers in the United States and Japan," *Frontiers in Psychology,* vol. Volume 12 - 2021, 2022.
[31] C. Longoni, and L. Cian, "Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The "Word-of-Machine" Effect," *Journal of Marketing,* vol. 86, no. 1, pp. 91-108, 2022.
[32] E. Aharoni, S. Fernandes, D. J. Brady, C. Alexander, M. Criner, K. Queen, J. Rando, E. Nahmias, and V. Crespo, "Attributions toward artificial agents in a modified Moral Turing Test," *Scientific Reports,* vol. 14, no. 1, pp. 8458, 2024.
[33] D. Ovsyannikova, V. O. de Mello, and M. Inzlicht, "Third-party evaluators perceive AI as more compassionate than expert humans," *Communications Psychology,* vol. 3, no. 1, pp. 4, 2025.
[34] H. Yamamoto, and T. Suzuki, "Exploring condition in which people accept AI over human judgements on justified defection," *Scientific Reports,* vol. 15, no. 1, pp. 3339, 2025.
[35] X. Ye, W. Jo, A. Ali, S. C. Bhatti, C. Esterwood, H. A. Kassie, and L. P. Robert, "Autonomy Acceptance Model (AAM): The Role of Autonomy and Risk in Security Robot Acceptance," in Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, Boulder, CO, USA, pp. 840–849, 2024.
[36] H. C. Triandis, R. Bontempo, M. J. Villareal, M. Asai, and N. Lucca, "Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships," *Journal of personality and Social Psychology,* vol. 54, no. 2, pp. 323, 1988.

This is an Accepted Manuscript of an article published by Taylor & Francis in Advanced Robotics on 25 Sep 2025, available at: https://doi.org/10.1080/01691864.2025.2561627

[37] S. Yamamoto, "The Social Possibilities of Economic Bushido: Inazo Nitobe's Bushido: The Soul of Japan and its Application to Modern Society," 2019.

[38] N. Dreksler, H. Law, C. Ahn, D. Schiff, K. J. Schiff, and Z. Peskowitz, "What Does the Public Think About AI? An Overview of the Public's Attitudes Towards AI and a Resource for Future Research," 2025.

[39] A. Javaheri, N. Moghadamnejad, H. Keshavarz, E. Javaheri, C. Dobbins, E. Momeni-Ortner, and R. Rawassizadeh, "Public vs media opinion on robots and their evolution over recent years," *CCF Transactions on Pervasive Computing and Interaction,* vol. 2, no. 3, pp. 189-205, 2020.

[40] Y. Huang, E. Richter, T. Kleickmann, and D. Richter, "Comparing video and virtual reality as tools for fostering interest and self-efficacy in classroom management: Results of a pre-registered experiment," *British Journal of Educational Technology,* vol. 54, no. 2, pp. 467-488, 2023.

[41] Q. Xu, J. Ng, O. Tan, Z. Huang, B. Tay, and T. Park, "Methodological Issues in Scenario-Based Evaluation of Human–Robot Interaction," *International Journal of Social Robotics,* vol. 7, no. 2, pp. 279-291, 2015.

[42] M. Mara, J.-P. Stein, M. E. Latoschik, B. Lugrin, C. Schreiner, R. Hostettler, and M. Appel, "User Responses to a Humanoid Robot Observed in Real Life, Virtual Reality, 3D and 2D," *Frontiers in Psychology,* vol. Volume 12 - 2021, 2021.

[43] N. Randall, and S. Sabanovic, "A Picture Might Be Worth a Thousand Words, But It's Not Always Enough to Evaluate Robots," in Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden, pp. 437–445, 2023.