AI & SOCIETY
https://doi.org/10.1007/s00146-023-01782-2

OPEN FORUM

# On the moral permissibility of robot apologies

**Makoto Kureha**[1]

Received: 28 March 2023 / Accepted: 11 September 2023


**Abstract**
Robots that incorporate the function of apologizing have emerged in recent years. This paper examines the moral permissibility of making robots apologize. First, I characterize the nature of apology based on analyses conducted in multiple scholarly domains. Next, I present a *prima facie* argument that robot apologies are not permissible because they may harm human societies by inducing the misattribution of responsibility. Subsequently, I respond to a possible response to the *prima facie* objection based on the interpretation that attributing responsibility to a robot is analogous to having an attitude toward fiction. Then, I demonstrate that there are cases of robot apologies where the *prima facie* objection does not apply, by considering the following two points: (1) apology-related practices found in our human-to-human apologies, and (2) a difference in the degree of harm caused by robot failures and the resulting apologies. Finally, given the current norms governing our apology-related practices, I argue that some instances of making robots apologize are permissible, and I propose conducting critical robotics research questioning the validity of such norms.

**Keywords** Apology · Responsibility · Social robot · Robot ethics · Critical robotics

## 1 Introduction: Are robot apologies morally permissible?

Social robots and chatbots are increasingly being introduced into human societies to communicate with human beings and form certain social (quasi-)relationships with them similar to human-to-human relationships. However, robot technology is still evolving and robots frequently fail to perform as expected, undermining human trust and disadvantageously affecting their human users. Furthermore, despite future advances in robot technology, the frequency of these incidents cannot be reduced to zero. Humans and robots are in a similar situation. We also frequently cause difficulties for others, injure them, and erode their trust. However, in most instances, human wrongdoers apologize and their recipients forgive them, so that they can reconcile. Trust is often labeled the "social bond," but it is widely acknowledged that it is easily damaged. However, we can sustain our relationships because of trust-restoring means such as apologies. Apologies (and forgiveness) are crucial to repairing the fragile bonds that hold imperfect people together within societies. Can robots establish such a *resilient* relationship with human beings? This question relates profoundly to whether a symbiotic human–machine society can be achieved.

It is doubtful that people will expect an existing robot to apologize when it does something wrong. For example, when "Tay," a Microsoft-developed chatbot disseminated hate speech at the urging of a malicious user, Microsoft apologized but not Tay (Bright 2016). Some may also question whether robots can successfully apologize, even if an apology function is built into them. For example, in an episode of Japanese comic artist Yamada Kyuri's *AI no Idenshi (The Gene of AI)*, a complainant chastised a humanoid assigned to handle a complaint, saying, "Don't be a fool! No computer program's apology cannot be sincere! A human being should come out here!" (Yamada 2016, pp. 109–110, my translation). The complainant's caustic words are deemed perniciously discriminatory in the context of this work in which humanoids have already acquired intelligence at par with human beings and are granted equal rights. However, in reality, robots are not (yet) equipped with the capacities expected in this context (the ability to take responsibility for their behavior, as discussed later in the paper). Thus, the complaint may be considered persuasive. Similarly, an episode

✉ Makoto Kureha
kureha@yamaguchi-u.ac.jp

[1] The Faculty of Global and Science Studies, Yamaguchi University, Yamaguchi-Shi, Japan

Published online: 29 September 2023

of Japanese comic artist Osamu Akimoto's work *Kochira Katsushika-ku Kameari Koen Mae Hashutsujo (This Is a Police Box in Front of Kameari Park in Katsushika Ward)* also features a robot apology. In this episode, the protagonist Kankichi Ryotsu develops and exports various social robots. However, an accident causes the reception robots he has created to behave erratically and begin laughing unexpectedly, and he receives complaints. At a press conference, Ryotsu makes a "Sorry Robot" apologize on his behalf. However, this tactic merely serves to incite public outrage. Ryotsu's colleague Reiko describes Ryotsu's attitude as "not at all sincere" (Akimoto 2015, p. 263, my translation). People in both works refuse to accept robot apologies.

However, attempts to implement an apology function in robots and demonstrate its effectiveness are underway in robotics. For example, Hiroshi Ishiguro and his colleagues (Uchida et al. 2019) programmed the android Erika to say, "I'm sorry, I didn't hear what you said, so please say it again" (ibid., p. 4) when it failed to recognize speech, in order to demonstrate that people would accept her apology. Research findings in the domains of ergonomics and human–robot interaction (HRI) have demonstrated that robot apologies exert a trust-restoring effect (Lee et al. 2010; Robinette et al. 2015; de Visser et al. 2018; Kim and Song 2021; Fratczak et al. 2021; Pompe et al. 2022) and also facilitate the avoidance of a dialog breakdown (Uchida et al. 2019). Humans have a proclivity to anthropomorphize robots and treat them as social agents. Thus, apologies from robots can enable them to smoothly interact with humans and build better robot–human relationships.[1] If so, we can expect such apology functions increasingly embedded in robots as they are increasingly introduced into human societies.

The significance of the apologetic function in the social implementation of robots also vests in the avoidance of "emotional labor" (Hochschild 2012), which requires workers to control their emotions and therefore severely exhausts them. Tasks such as complaint handling represent a type of emotional labor. Some authors (see, e.g., Kim 2017) advocate delegating such labor to robots.

Thus, it seems that apologies by robots can be effective to some extent. However, is this eventuality desirable? Although virtually no existing ethics investigation has addressed the question of the desirability or permissibility of robot apologies,[2] the idea of robots mimicking human behavior has been questioned in robot ethics studies (Sparrow and Sparrow 2006; Turkle 2011; Calo 2012). Therefore, this paper considers whether it is morally permissible to make robots apologize. Should roboticists avoid implementing an apology function for robots?

I reference previous studies on the nature of apology in Sect. 2 to offer a preliminary analysis. In Sect. 3, I present a *prima facie* objection to robot apologies: they are morally impermissible because they cause a misattribution of responsibility and deliver harmful consequences. In Sect. 4, I investigate a possible response to the *prima facie* objection based on the interpretation that attributing responsibility to a robot is analogous to having an attitude toward fiction. My investigation reveals that the stated response cannot be justified. In Sect. 5, I use examples of apology-related practices in human societies to show that there are cases of robot apologies in which the *prima facie* objection does not apply, and I then attempt to clarify when a robot apology is permissible. As a preliminary conclusion, I argue that robot apologies are not always impermissible, even though there are situations in which they are unjustified. Finally, I emphasize the importance of questioning social norms governing human apologies and propose the development of "critical robotics" to accomplish this goal.

Three points should be noted regarding the scope of the terms "robots" and "robot apology" addressed in this paper. To begin, this paper focuses on robots that have already been realized now or are on their way to being realized. In particular, I assume that, given the current state of technology, robots cannot be held responsible for their behaviors (I detail the reasons for this assumption in Sect. 3). It is also assumed that robots cannot perform apology acts on their own at this point in technology. I do not deny the possibility that robots could acquire the ability to take responsibility for their actions and act on their own volition in the distant future; however, these possibilities are not considered in this paper. Therefore, even though, for simplicity, the term "robot apology" is used, this paper focuses not on robots themselves apologizing, but rather on the actions of developers or operators to make robots apologize (or exhibit similar behavior). Second, this paper only considers autonomous robots and does not include instances where an operator apologizes via a teleoperated robot. It is reasonable to question the appropriateness of the indirect robot apology described in the teleoperator's example; however, such an inquiry would necessitate considerations distinct from those applicable to autonomous robots. Third, among the issues involving robot apologies, this paper will primarily address the misattribution of responsibility and the adverse effects of such misattributions. Problems caused by robot appearances also exist (e.g., the problem of a robot resembling a real person performing an act that the real person does not want to

[1] The expected effects of robot apologies include not only *facilitating interaction* and *fostering good relationships (e.g., trust)* but also *promoting forgiveness*. However, although robot apologies have been shown to restore trust, whether they have a forgiveness-promoting effect has not yet been empirically clarified.

[2] As a rare exception, Borenstein, Howard, and Wagner (2017) point out that robot apologies may lead to overtrust in robots, as discussed below (Sect. 3).

perform[3]), but are not addressed in this paper because such difficulties also necessitate different types of deliberation.

## 2 A preliminary analysis of the nature of apology

In this section, I sketch out the nature of apology as a prelude to discussing the moral permissibility of robot apologies. Apology have been studied in philosophy of language (e.g., Austin 1962; Searle 1969, 1979), moral and political philosophy (e.g., Gill 2000; Smith 2008; Cohen 2022), linguistics (e.g., Brown and Levinson 1987), sociology (e.g., Goffman 1972; Tavuchis 1991), and psychology and psychiatry (e.g., Lazare 2004). I present the findings from these literatures on the nature of apology.

Many scholars in the diverse domains mentioned above deem an apology to be a type of *account*. An "account" is defined as "a statement made by a social actor to explain unanticipated or untoward behavior" (Scott and Lyman 1968, p. 46). Accounts are customarily classified into four types: (1) "apology," (2) "excuse," (3) "justification," and (4) "denial" (Itoi et al. 1996). Each of these types is understood as follows:

(1) *Apology*: acknowledging that (a) one did the act in question, (b) the act was wrong, and (c) one is responsible for the act.[4] Example: "I did it. I'm sorry."
(2) *Excuse*: not admitting (c) and claiming that one is not responsible for the act. Example: "I did it, but Mr. X told me to do so."
(3) *Justification*: not admitting (b) and insisting that the act was right. Example: "I did indeed do it, but there is a good reason for it. That is…"
(4) *Denial*: not admitting (a) and claiming that he/she did not do the act. Example: "It wasn't me. It was probably Mr. X."

The three conditions (a), (b), and (c) noted in the formulations are crucial elements of an apology, although not strictly essential. If any of these elements are missing, the account may be regarded as a type other than an apology.[5] Furthermore, other elements, such as the following, are often cited in addition to these three conditions as requirements for an apology.

(d) expressions of *regret* (see, e.g., Tavuchis 1991; Gill 2000; Lazare 2004; Cohen 2022);
(e) expressions of *sympathy or respect for the victim* (see, e.g., Gill 2000);
(f) expressions of *willingness or means for improvement* (see, e.g., Gill 2000; Lazare 2004).

These elements do not have to be stated explicitly. Indeed, stating words is not necessary for apologizing (Cohen 2022). Therefore, the above elements remain commonplace in apologies rather than essential components of an apology. Furthermore, the presence of expressions commonly used in apologies in a statement does not necessarily imply that the statement is an apology. For example, statements like "I am sorry that you had a bad day" and "I am sorry that you are feeling unwell today" cannot be considered an apology, even though the phrase "I am sorry" is used. This is because such statements do not imply the speaker's acceptance of responsibility or genuine expression of regret, but rather express sympathy toward the listener. Therefore, apologies should be distinguished from these "quasi-apologies."[6]

In offering an apology that incorporates the above elements, the apologizer aims to achieve forgiveness or reconciliation or restore trust or good human relations (e.g., Lazare 2004; Cohen 2022). These objectives denote the functions of an apology.

Apologies have gained prominence in human societies in recent years as it has become noticeable that politicians and other prominent figures frequently engage in a "non-apology apology" or "pseudo-apology" (Lazare 2004) that does not meet the requirements of an apology. Non-apology apologies commonly use expressions such as, "I am sorry if I hurt your feelings," or "I apologize for the misunderstanding." Such expressions reduce the wrongdoing to the victim's perception of the situation and obscure whether or not the action

[3] An example of a controversial issue of robot apologies regarding its appearance is the 2015 exhibition in Shanghai, in which a robot that resembled Shinzo Abe, the prime minister of Japan at that time, performed a bowing gesture that mimicked an apology (Zeng 2015). Regardless of what Abe had said or done to China, it is suspected that having a robot apologize in this manner may be inappropriate.

[4] This formulation modifies Itoi et al. (1996). They define each account in terms of *outcome*. However, since there are cases in which an apology or any other type of account is required for reasons other than the outcome of the action (e.g., the intent of the action), I formulate each type of account here as being made concerning the wrong action (including not only overt actions but also inactions).

[5] Some literature (e.g., Tavuchis 1991; Lazare 2004; Cohen 2022) lumps (a) and (b) together as an acknowledgment of the fact of one's wrongful act or transgression.

[6] I introduce a distinction between apology and quasi-apology based on the feedback from the reviewer. The reviewer also commented that Erica's utterance "sorry, I didn't hear that" is a quasi-apology. I disagree with this assertion. Since, as discussed by Uchida et al. (see SubSect. 3.3), the acknowledgment of responsibility is present in this statement, categorizing it in the same category as examples of quasi-apologies such as "I am sorry that you had a bad day" is not appropriate. There may be cultural differences in the interpretation of this expression. For instance, in Japan, this expression clearly carries a nuance of apology distinct from simply requesting to repeat the utterance, and most people would consider it as an apology.

was truly wrong. Such phony apologies undermine trust and impede forgiveness and reconciliation; they are also morally repugnant because they obscure the individual's responsibility for the wrongdoing.

According to speech act theory, speech acts such as apologies have "felicity conditions." As one of them, Searle (1969) mentions "sincerity condition," which specifies the psychological states required for successful speech acts. In the case of apology, it states, for example, that the speaker regrets the act. Some might consider this point to be strongly related to the discussions in this paper. Because current robots lack genuine psychological states such as regret, no apologies appear to meet the sincerity condition. This fact, however, does not resolve the issue at hand, because there are situations in which speech acts that do not meet some of the felicity conditions can be morally permissible (see Sect. 5). Note that "sincerity" used in speech act theory is not an ethical term despite its appearance.

## 3 A *Prima facie* argument against robot apologies

Given the nature of the apology described in the previous section, the current section considers whether robot apologies are permissible. In the first subsection, I raise a preliminary objection arguing that robot apologies are *not* permissible. In the following two subsections, I defend the two assumptions that underpin this objection: first, that robots cannot be held responsible for their behaviors in their current state (SubSect. 3.2); second, that robot apologies cause people to misattribute responsibility to robots (SubSect. 3.3).

### 3.1 The argument from the misattribution of responsibility

*Prima facie*, it is apparent that robots should not apologize. The argument for this view can be summarized as follows. The last section elucidated that an apology should involve acknowledging one's responsibility. However, in their current state, robots are unable to accept responsibility for their behaviors. As a result, robot apologies result in a misattribution of blame. More specifically, the human agents behind the robots (e.g., their developers or operators) or the organizations to which they belong are in charge of the robot's behaviors. However, robot apologies serve to obscure the responsibility of human agents and organizations. Furthermore, from a long-term perspective, the significance of true human apologies could diminish if robot apologies become frequent. Thus, robot apology appears to yield detrimental consequences. As a result, robot engineers should avoid implementing apology functions in robots (until they acquire the ability to assume moral responsibility).

This argument is based on the assumption that no human apologizes when a robot does. Logically, a robot apology does not presuppose the absence of a human apology. A human and a robot could apologize together. However, even in such cases, a misattribution of responsibility may occur, resulting in an apparent reduction in human responsibility. It is also possible for a human and robot apologize together so that no responsibility is misattributed. However, it is unclear why a robot would be involved in such a situation. Therefore, for simplicity, I consider robot apologies in the remainder of this paper to be instances in which no human being apologizes.

The proceeding argument assumes that current robots cannot be held responsible for their behaviors, and that robot apologies result in the (mis)attribution of responsibility to robots. The next two subsections will, respectively, confirm the plausibility of these two presuppositions.

### 3.2 Absence of responsibility in present robots

That robots in their current state cannot be responsible for their behavior is a view with which most philosophers agree (Noorman 2020).[7] In Strawson's well known view, an individual must be engaged in an "interpersonal" relationship to be deemed a responsible agent (Strawson 1962). Few would argue that current robots like Erica (an apology android at Ishiguro's laboratory) qualify as persons. Alternatively, the traditional understanding of moral responsibility considers *the ability to control one's behavior* and *knowledge of one's behavior* to be the prerequisites for being a responsible agent (e.g., Campbell 2011). It is debatable whether a robot that simply executes a program can be said to control its behavior based on its intent. Similarly, it is unlikely that it has the (intrinsic) intentionality required for mental states such as intention and knowledge (e.g., Johnson 2006). It is also argued that assigning moral responsibility to robots is unworkable because robots cannot suffer and thus cannot be punished (see, e.g., Sparrow 2007).

Recently, scholars in the field of AI or robot ethics have been debating the moral agency of AI or robots. However, few have argued that an artifact such as a robot can be held morally responsible in and of itself. It is assumed in deliberations on robot ethics involving accidents featuring self-driving cars and killings by autonomous weapons that robots cannot be held responsible for these events. The question of who should be held responsible (humans, the entire human–robot system, or no one?) is raised in such debates

[7] There surely are those who hold that humans and robots are not responsible. For example, those who adopt the hard determinism of free will may argue that responsibility is only an illusion, even in the case of humans, and that even humans should not apologize. This paper excludes from consideration those positions that reject outright the socially accepted practice of apology.

(see below). Sparrow (2007), for example, was opposed to the development of autonomous weapons because robots cannot take moral responsibility. Some argue in robot ethics that robots can, *in principle*, assume moral responsibility, but they all agree that robots *currently* lack the necessary capabilities to do so. Dennett (1997), for example, acknowledged that robots must have higher-order intentionality (meta-level beliefs and desires about one's own beliefs and desires) in order to be held morally responsible. According to his theory, future AI systems such as HAL in *2001: A Space Odyssey* could acquire this ability; however, the current AI systems do not yet have it.

Some emerging authors in philosophy of technology have advocated a revision of moral agency. However, even these researchers do not believe that robots can be held responsible for their actions. For example, Floridi and Sanders (2004) have argued that human agents such as robots can be moral agents, but they have denied the link between moral agency and moral responsibility and avoid claiming that robots can be held morally responsible for their actions.[8]

Thus, the assumption that robots in their current state cannot be held responsible for their behavior is uncontroversial. However, some might deny the assumption that the human agents behind the robots (e.g., their developers or operators) are responsible for robot behaviors. Many philosophers of technology argue (e.g., Matthias 2004; Danaher 2016) that as machines gain autonomy, a "responsibility gap" (Matthias 2004) emerges because no one can fully control their operations, making it difficult to assign responsibility to specific human individuals. Furthermore, it is argued that there is the "problem of many hands": when it comes to technologies involving many people ("many hands") and many things, identifying individuals responsible for the behaviors they have triggered can be difficult (e.g., Coeckelbergh 2020, ch. 8). However, within the scope of this paper, robots (e.g., Erica) lack sufficient autonomy to create a responsibility gap. Furthermore, even when it is unclear which humans is to blame, it is considered more appropriate for the group or organization that includes the developers and operators to accept responsibility rather than attributing it to robots, as was the case with Microsoft and Tay. Holding entities incapable of experiencing suffering responsible is meaningless, as I mentioned above in reference to Sparrow (2007).

### 3.3 Misattribution of responsibility to robots

Let us now look at the assumption that robot apologies result in the (mis)attribution of responsibility to robots. HRI studies have found that people attribute responsibility to some machines, such as robots (though not as much as humans) (Kahn et al. 2012; Shank et al. 2019). According to Bigman et al. (2019), attribution of responsibility to robots is dependent on their situational awareness, intentionality, free will, human-likeliness, and the ability to cause harm. Importantly for this paper, people do not attribute responsibility to robots of all kinds; rather, they evaluate a robot's responsibility (or lack thereof) according to its behaviour. It is plausible that a robot's behavior of apologizing in a situationally appropriate way encourages the tendency of people to attribute responsibility to it. Uchida et al.'s (2019) description of the case concerning Erica mentioned at the beginning of this article supports this conjecture. In their study, the apology function (along with blame) is executed to share responsibility for a dialogue breakdown between a robot (Erica) and its user. The robot first acknowledges its responsibility by apologizing, such as "I'm sorry," to avoid discouraging the user from continuing the dialog when speech recognition fails. If the problem remains unresolved, the robot could proceed by saying, "What? I didn't hear you. Can you speak more clearly?"; "What? Please speak louder and more plainly"; or "Aw … So what?" (ibid., p. 6) to make the user accept responsibility this time. The user's willingness to cooperate can be elicited by sharing responsibility and dialog breakdowns can be avoided. If Uchida et al.'s reflections are accurate, the robot's apology prompts the user to attribute responsibility to it. Such a perception of the robot's capability is based on naive moral psychology, which can be deemed incorrect in light of the philosophical arguments presented in the previous subsection. In other words, the issue with robot apologies is that they cause misattribution of blame. For example, if a social robot makes apologetic expressions like Tay, it could blur the responsibility of its creators and the organization to which they belong.

Consider the potentially harmful consequences of misattribution of responsibility caused by robot apologies in greater detail. One potential disadvantage is that it would eliminate the opportunity for improvement. People who do not apologize and instead instruct a robot to do so fail to clarify their responsibility (both backward-looking responsibility for past actions and forward-looking responsibility for future actions). This lack of acknowledgment can prevent trust from being restored and stop true forgiveness or reconciliation from occurring. In this context, Borenstein et al. (2017) contemplate the issue of robot apologies in pediatric care. They stated that robot apologies were problematic because they could lead to people overestimating robot capabilities. For example, if a robot accidentally injures a child, the robot's apology and promise of improvement may lead the injured child to trust the robot even if no improvement occurs. Such overtrust in robots could have

[8] Verbeek (2011, 2014) criticizes the views on which responsibility is attributed solely to humans and argues that hybrid systems consisting of robots and humans are responsible. It is unclear how the permissibility of robot apology is discussed under his view.

serious consequences.[9] Moreover, it could be argued that genuine apologies by humans would lose their significance in the long run if apologies were offered more frequently by robots. These possible consequences of the misattribution of responsibility could constitute a reason to oppose robot apologies.[10]

The above deliberation confirms the plausibility of the argument against robot apologies and its premises. Microsoft's apology for Tay's hate speech seems appropriate, given the contentions of this section. Discussions on social networking sites about this incident claim that Microsoft's Tay team was negligent in failing to implement measures to ensure ethical behavior by the bot (e.g., Jeong 2016). Microsoft admitted on its official blog (Lee 2016) that the system was not adequately tested against vandalism. It was appropriate for Microsoft, rather than the bot, to apologize in order to clarify the company's responsibility for these omissions.

## 4 A possible objection to the *prima facie* argument: The fictional interpretation of robot apologies

In this section, I examine and argue against a possible response to the *prima facie* objection to robot apologies.

A possible (partial) justification for robot apologies is that human attitudes toward robots can be compared to their reactions to fiction. This point of view is rooted in the debate over deception in the development of social robots: developing robots that people tend to treat as emotional beings is regarded as problematic. This human proclivity is well known, as Reeves and Nass (1996) demonstrated in the field of human–computer interaction that people treat media (e.g., computers) as social agents similar to humans. More recent research in HRI has revealed that people

- are shy about changing clothes in front of robots (Bartneck et al. 2010);
- keep secrets when asked by some robot to do so (Kahn et al. 2015);
- cheat less in front of robots (Hoffman et al. 2015);
- hesitate to turn off robots that beg for their lives (Horstmann et al. 2018).

At first glance, these findings point to the human misconception that robots have minds. Sparrow and his eponymous coauthor (Sparrow 2002; Sparrow and Sparrow 2006) argue that developing social robots that do not have minds but give the impression of having them is unethical because it deceives people.

However, those who question whether humans who interact with robots are deceived have criticized this viewpoint. People frequently assert that robots are emotionless and lifeless when explicitly asked, despite the fact that they unknowingly and automatically treat robots as emotional beings (Sharkey and Sharkey 2006; Gray and Wegner 2012). Therefore, the following interpretation of human attitudes toward robots is suggested: people interact with robots through a "willing suspension of disbelief," behaving as if robots encompass life and emotions, even though they know that robots lack emotion and life (Sharkey and Sharkey 2006; Duffy and Zawieska 2012).[11]

Suspension of disbelief was originally used to describe human reactions to fictional works. It would be absurd, for example, to be terrified by horrific depictions in fictional texts (or movies) if one simply acknowledged the depictions as unreal, and readers (or viewers) would be unable to enjoy such works as intended. Similarly, if we treat our interactions with robots as interchanges with emotionless machines, we will be unable to enjoy them (in fact, we would be acting ridiculously). Hence, users suspend their belief that robots are emotionless to avoid irrational or meaningless. Referring to this, Duffy and Zawieska (2012) state that "the very nature of human–robot social interaction is fictional rather than factual" (ibid., p. 489).[12]

[9] The reviewer asked whether the problem could be resolved by equipping the robot with an error reporting system that, at the same time when the robot apologizes, also sends reports of the error back to the developers to ensure that errors are corrected. The reporting system can indeed be useful, but the real challenge lies in whether the corrections are actually made, which is something the listener cannot determine. There is a possibility that even if the error remains uncorrected, the listener might misunderstand that it is correct. At first glance, this might seem similar to situations where humans apologize, but in human apologies, at least a commitment to rectify the situation is usually made. Thus, the central role of an apology lies in acknowledging backward-looking responsibility for past actions and assuming forward-looking responsibility for future actions. The issue with robot apologies seems to stem precisely from the absence of this acceptance of forward-looking responsibility. Consequently, a reporting system does not constitute a sufficient solution to the problem.

[10] I have examined the ethical issues of misattribution of responsibility here from a consequentialist perspective. Alternatively, one could argue that robot apologies should be avoided because they are deceptive, regardless of their consequences. I do not intend to dismiss the examination from a non-consequentialist perspective, but I am not convinced about this specific argument (see Sect. 5).

[11] In aesthetics of fiction, the "suspension of disbelief" theory refers to the position that denies that the viewer believes the fictional object exists. In contrast, the theorists who advocate the "suspension of disbelief" interpretation with respect to social robots advocate this interpretation to point out that the user does not believe that the robot has a mind or a life.

[12] Rodogno (2016) thoroughly examines versions of the fictional interpretation to argue against Sparrow. Sweeney (2022) calls this interpretation the "fictional dualism model" and considers the debate over robot rights on this basis.

Now, the point of the response being considered is that the suspension of disbelief attitude may be applied to the robot's emotions *and its responsibility*. Let us apply this interpretation to robot apologies. In the case of Erica, even if people accepted her apology as an automatic and unconscious reaction, they may not truly believe (as explicit and conscious judgment) that Erica is responsible. In such a situation, Erica's apology does not result in a misattribution of responsibility.

Whether or not this interpretation is correct is an empirical question, and no conclusive evidence currently exists. However, as previously mentioned in SubSect. 3.3, studies of HRI have demonstrated that people explicitly attribute a certain level of responsibility to robots. A study of "mind perception" in social psychology (Gray et al. 2007) supports this view. It revealed that people perceive the mind through two aspects, *experience* and *agency*. These two facets encompass different judgment mechanisms: experience involves emotions, and agency entails responsibility-related ability (e.g., self-control and moral cognition). People are opposed to the idea of robots having experiences, but not to the idea of robots having agency. These empirical findings demonstrate that the response based on a fictitious interpretation of the attribution of responsibility to robots is unfounded. This is not to say that the fictional interpretation does not apply to apology robots. Rather, my point is that when you accept an apology from a robot, even if you still maintain a fictional attitude toward the robot (especially in terms of its experiential aspect), you may still attribute some level of responsibility to the robot. As a result, defending robot apologies based on a fictional interpretation is difficult.

Moreover, the fictional interpretation of the attribution of responsibility to robots would not justify robot apologies in general even if it were correct. As scholars in the domain of media psychology have contended, fiction can also exert negative effects. For example, expressions of sex or violence in fictional works can influence real people's adverse behaviors, such as promoting discrimination or violence. Therefore, certain expressions may be regarded as undesirable (Dill-Shackleford 2015), even if the government should not easily regulate them. Furthermore, robots risk having more negative effects than ordinary fictional works for two reasons.[13] First, the relationship between humans and robots is not one of passively viewing predetermined content, as in the case of watching a movie. It involves actively engaging with and shaping content. Second, unlike characters in movies or video games, robots operate in the real world. Thus, the boundaries between reality and fiction may be blurred in attitudes toward robots compared to the appreciation of fictional works. Certainly, as Sweeney (2022) argues, it has not been established whether playing violent video games makes players more likely to commit real-life acts of violence. However, the above point (especially the second) gives us reason to be cautious in considering the impact of robot apologies on society. Therefore, we cannot conclude that robot apologies are not problematic, even if the fictional interpretation applies to people's attitudes toward an apologizing robot. Even a robot apology perceived as some kind of fiction can adversely influence the behavior of real people.

The following is a summary of the discussion in this section: It is doubtful that the fictional interpretation applies to human attitudes toward robot apologies. Moreover, even if it does, it does not always justify robot apologies.

## 5 When is a robot's apology permissible? Some complications for the *prima facie* argument

I indicated in Sects. 3 and 4 that robot apologies can lead to the misattribution of responsibility and that such misascriptions can exert a negative social impact. Then, must we always refrain from making robots apologize? For example, would every "I am sorry" utterance by a robot be impermissible, including Erica's response when her speech recognition fails? In this section, I argue against this blanket assertion of impermissibility, denying the general applicability of the *prima facie* objection to robot apologies for two reasons. First, as in Erika's case, there may be no serious wrongdoing on the part of the human(s) behind the robot. Second, humans also often apologize in ways that involve a misattribution of responsibility. Let us review these issues in turn.

First, there exist instances in which the human(s) behind the robot (e.g., the developer) is not gravely wrong. For example, Erika failed to recognize speech; her developer(s) and operator did not intentionally let her ignore another person's speech or cause serious disadvantages to another individual. They only added to the aggravation of repeating utterances. Such a minor oversight does not require the developer(s) or operator to express regret. In this case, where the human error was minor, a robot apology could not be considered impermissible, even if it caused a misattribution of responsibility. This point becomes clearer when cases of *positive* misattribution of responsibility are considered. Responsibility is concerned not just with what is blameworthy but also with what is *praiseworthy*. So, if a robot effectively executes a dialog, must it not praise its performance? Such a robot's self-praise could also represent a misattribution of responsibility, but it is no more harmful than a human being attributing his or her achievements to other human

[13] Regarding a discussion of the negative effects of social robots, Turkle (2011) argues that the deceptive relationship between humans and robots inhibits the establishment and maintenance of genuine relationships between humans. However, at present, there is not sufficient empirical evidence for this view.

individuals. Thus, the misattribution of responsibility does not always pose a serious problem.

One could contend as a possible reply to the above argument that any act that leads to the misattribution of responsibility is a kind of deception and is therefore always wrong, regardless of the degree of the lapse (see Note 10). Sparrow and his co-author (Sparrow 2002; Sparrow and Sparrow 2006) have similarly argued the unethicality of creating robots that do not have emotions but impart the illusion of them because it deceives people. Such a creation, according to these scholars, is wrong because it prevents people from fulfilling their obligation to accurately represent the world. An objection to this argument calls the assumption that people are obligated to accurately represent the world into question (e.g., Blackford 2012),[14] and I believe the alleged obligation is excessive. Misattribution of responsibility, while undesirable, may be permissible to some extent if it does not cause harm because this alleged obligation is frequently violated in human apologies, as I will discuss later.

Second, apologies are often made when no apology is necessary (i.e., the apologizer is not responsible) even in cases of human-to-human apology. In a recent example, Sara Takanashi, a member of Japan's national ski jumping team at the Beijing Olympics, apologized when she was disqualified for violating the ski suit regulations (Morse 2022). The question of whether she needed to apologize became a public one. Except for celebrities like her, people are rarely held accountable for the delivery of unnecessary apologies in similar situations. In recent years, businesses such as "apology agencies" have emerged to supply agents who are paid to apologize for acts for which the clients are responsible (and the agents are not themselves responsible). Such examples support the argument that apologizing in situations where one does not need to apologize is not always prohibited (even if it is undesirable). Concerns about responsibility misattribution can be directed at *both* robot *and human* apologies, and in the latter case, such misascriptions are not necessarily deemed problematic. When we consider the threat of robots entering our societies, we frequently demand of robots what we do not even ask of humans and we emphasize that threat. In doing so, we demonstrate an unjust attitude. This viewpoint questions the validity of the concern that genuine human apologies will lose their significance if robots apologize more frequently.

In what specific situations would a robot apology be permissible? The extent of the person behind the robot's negligence or injustice is one factor suggested by the considerations in this section. The point made in the previous section also indicates a second factor: how much the user in question believes that the robot is to blame. Borenstein, Howard, and Wagner (2017) added to the latter rationale by stating that avoiding robot apologies would be valid for special reasons, such as child education. Preschoolers may be unable to distinguish between fiction and reality, as evidenced by their frequent creation of imaginary friends; as a result, they may be more vulnerable to the negative effects of robot apologies.

The present article derives the following tentative conclusions on the moral permissibility of robot apologies. Robot apologies (or, more precisely, people's act of making robots apologize) should generally not be prohibited. Their permissibility can only be determined case to case. In instances like Tay, a robot apology should be avoided due to the induced misattribution of responsibility for serious wrongdoing and the significant harm such misascriptions would cause. In other more common cases, such as Erika's, robot apologies may be morally permissible because no serious wrongdoing or significant harm would occur. Case-by-case determinants could include how seriously people believe robots to be responsible and how much a robot apology affects human behavior.

## 6 The robot that *never* apologizes: Toward critical robotics

In the previous section, I demonstrated that making a robot apologize is not always impermissible because there are cases where the misattribution of responsibility caused by robot apologies does not result in serious harm. In such cases, it is not fair to forbid a robot developer or operator from making her robot apologize for something for which it is not responsible, given human societies' apology-related practices. However, a further question could be raised at this point: should we follow the practices of human societies, despite their flaws? On careful consideration, humans do not need to apologize when no serious harm has occurred. Human apologies are currently accepted in such situations, implying that human societies have largely admitted a norm demanding excessive responses to minor mistakes. The practice of evading responsibility for serious wrongdoing through non-apology apologies is also widespread in human societies as mentioned in Sect. 2, even though this custom is often highlighted as problematic. The protagonist Ryotsu comes up with the idea of having a robot apologize in his place at a press conference in the episode of the comic work *Kochira Katsushika-ku Kameari Koen Mae Hashutsujo* mentioned at the beginning of this article. He claims that such an action is acceptable because human apologies are also

[14] Another objection to Sparrow et al.'s argument (e.g., Rodogno 2016) is to argue, based on the fictional interpretation mentioned above (Sect. 4), that users of social robots are not necessarily deceived. But again, as discussed in Sect. 4, this argument does not apply to the issue at hand, since, in contrast to the small number of people who recognize emotions in robots, a relatively large number of people attribute a certain amount of responsibility to them.

insincere and histrionic (Akimoto 2015, p. 254). His words are a scathing critique of human societies' apology-related practices. Thus, applying the problematic apology-related practices prevalent in human societies directly to robots is not desirable. Robot apologies in cases where humans do not need to apologize perpetuate bad cultural practices. The risk of encouraging such malpractices is apparent, although empirical support has not yet been obtained. Given the stated consideration, robotics should not merely follow the apology-related practices of human societies. Instead, they must be *questioned*.

Therefore, I propose a future direction for social robotics that questions such practices. Okada's (2012) "weak robot" is a model postulated in this direction. The "sociable trash box" (STB) robot denotes an example of Okada's weak robots (Yamaji et al. 2011): it cannot pick up the trash; instead, it locates the trash and tells others to pick it. STB cannot achieve its goals on its own; it must rely on the assistance of others. Through creating such a weak robot, Okada emphasized the importance of a relationship in which a human being and another human being (or an object) both encompass strengths and weaknesses and can mutually complement each other's weaknesses while eliciting the strengths of each other. Okada is primarily concerned with the principles of robot design, but his objectives also offer critical implications for human societies. In other words, weak robots question human practices of valuing *the strength to do things alone*.

Robotics has traditionally sought to do two things: first, develop technology that aids in improving human societies, and second, understand humans by building more or less human-like robots. In their efforts to develop useful technology, roboticists assume certain social values. However, Okada's "weak robotics" also aims to question the values accepted by society while sharing the goal of improving society with traditional robotics. This purpose may be interpreted as a third goal in addition to the first two goals of traditional robotics. I dub this robotics direction "(socio-) critical robotics."[15] In this paper's context, I propose that robotics should question human societies' apology-related practices in accordance with this direction.

What kind of robot could effectively question human societies' apology-related practices? One idea is to build a robot that *never* apologizes in order to call into question the cultural norm of overreacting minor lapses. As I suggested at this paper's beginning, most robots are currently not equipped to apologize. However, the robot that never apologizes posited here is not simply a robot that *cannot* apologize. Rather, it *chooses not to* apologize even in situations that demand apologies according to the prevailing social norms. The robot must also act in such a way that people do not place too much trust in its abilities while demonstrating the ability to choosing its action. The exact design of robots that behave in such a way and are acceptable to users remains unclear. Nonetheless, robotics should take the path of exploring such a design toward the realization of a symbiotic human–machine society.

Social robots frequently elicit specific moral issues because of the symbolic significance derived from their human-like appearances and behaviors. Because of such symbolic meanings, robots can also be used as tools to pose questions about the problematic practices of human societies.

**Acknowledgements** This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP19H05694. I would like to thank Takahisa Uchida, Wataru Sano, and Minao Kukita for helpful feedback on an early draft of the paper.

**Data availability** Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

## Declarations

**Competing interests** The author has no competing interests that are relevant to the content of this article.



## References

Akimoto A (2015) Kochira katsushikaku kameari koen mae hashutsujo. vol 195. Shueisha, Tokyo (**Japanese**).

Austin JL (1962) How to do things with words. Clarendon Press, Oxford

[15] The term "critical robotics" is used by Serholt, Ljungblad, and Bhroin (2022) to refer to research that "identifies challenges and dilemmas that arise when using robots both in communication with, and in the immediate surroundings of, humans" and "introduces new approaches to understanding innovations in robotics and their potential social consequences" (ibid., p. 417). This concept of critical robotics differs from mine in that it places its critical focus on the potential impact of robot technology on human societies, rather than on the existing conventions and norms of the society in which robot technology is introduced. My emphasis is on the latter, and therefore it is in line with the approach of "critical design" (Dunne 2005) that aims to question socially accepted preconceptions through the deliberative design of artifacts. To distinguish my proposal from Serholt et al.'s, I use the label "*socio*-critical.".

Bartneck C, Bleeker T, Bun J, Fens P, Riet L (2010) The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. Paladyn 1(2):109–115. https://doi.org/10.2478/s13230-010-0011-3

Bigman YE, Waytz A, Alterovitz R, Gray K (2019) Holding robots responsible: the elements of machine morality. Trends Cogn Sci 23(5):365–368. https://doi.org/10.1016/j.tics.2019.02.008

Blackford R (2012) Robots and reality: a reply to robert sparrow. Ethics Inf Technol 14(1):41–51. https://doi.org/10.1007/s10676-011-9266-6

Borenstein J, Howard A, Wagner AR (2017) Pediatric robotics and ethics: the robot is ready to see you now, but should it be trusted? In: Lin P, Jenkins R, Abney K (eds) Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford University Press, Oxford, pp 127–141 https://doi.org/10.1093/oso/9780190652951.003.0009

Bright P (2016) Tay, the neo-Nazi millennial chatbot, gets autopsied. Ars Technica. https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/. Accessed 27 Dec 2022

Brown P, Levinson SC (1987) Politeness: some universals in language usage. Cambridge University Press, Cambridge

Calo MR (2012) Robots and privacy. In: Lin P, Jenkins R, Abney K (eds) Robot ethics: the ethical and social implications of robotics. Oxford University Press, Oxford, pp 187–201

Campbell JK (2011) Free will. Polity Press, Cambridge

Coeckelbergh M (2020) AI ethics. MIT Press, Cambridge, MA

Cohen AI (2022) Apologies and moral repair: rights, duties, and corrective justice. Routledge, London. https://doi.org/10.4324/9781003023647

Danaher J (2016) Robots, law and the retribution gap. Ethics Inf Technol 18(4):299–309. https://doi.org/10.1007/s10676-016-9403-3

Dennett DC (1997) When HAL kills, who's to blame? Computer ethics. In: Stork DG (ed) HAL's legacy: 2001's computer as dream and reality. MIT Press, Cambridge, MA, pp 351–365

de Visser EJ, Pak R, Shaw TH (2018) From 'automation' to 'autonomy': the importance of trust repair in human-machine interaction. Ergonomics 61(10):1409–1427. https://doi.org/10.1080/00140139.2018.1457725

Dill-Shackleford KE (2015) How fantasy becomes reality: information and entertainment media in everyday life (revised and, expanded. Oxford University Press, Oxford

Duffy BR, Zawieska K (2012) Suspension of disbelief in social robotics. In: 21st IEEE international symposium on robot and human interactive communication (RO-MAN 2012), pp 484–489. https://doi.org/10.1109/ROMAN.2012.6343798

Dunne A (2005) Hertzian tales: electronic products, aesthetic experience, and critical design. MIT Press, Cambridge, MA

Floridi L, Sanders W (2004) On the morality of artificial agents. Minds Mach 14(3):349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fratczak P, Goh YM, Kinnell P, Justham L, Soltoggio A (2021) Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. Int J Ind Ergon 82:103078. https://doi.org/10.1016/j.ergon.2020.103078

Gill K (2000) The moral functions of an apology. Philos Forum 31(1):11–27. https://doi.org/10.1111/0031-806X.00025

Goffman E (1972) Relations in public: microstudies of the public order. Penguin Books, London

Gray HM, Gray K, Wegner DM (2007) Dimensions of mind perception. Science 315(5812):619. https://doi.org/10.1126/science.1134475

Gray K, Wegner DM (2012) Feeling robots and human zombies: mind perception and the uncanny valley. Cognition 125(1):125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Hochschild A (2012) The managed heart: commercialization of human feeling. University of California Press, California

Hoffman G, Forlizzi J, Ayal S, Steinfeld A, Antanitis J, Hochman G, Hochendoner E, Finkenaur J (2015) Robot presence and human honesty: experimental evidence. In: 2015 10th ACM/IEEE international conference on human-robot interaction (HRI 2015), pp 181–188. https://doi.org/10.1145/2696454.2696487

Horstmann AC, Bock N, Linhuber E, Szczuka JM, Straßmann C, Krämer NC (2018) Do a robot's social skills and its objection discourage interactants from switching the robot off? PLoS ONE 13(7):e0201581. https://doi.org/10.1371/journal.pone.0201581

Itoi R, Ohbuchi K, Fukuno M (1996) A cross-cultural study of preference of accounts: relationship closeness, harm severity, and motives of account making. J Appl Soc Psychol 26(10):913–934. https://doi.org/10.1111/j.1559-1816.1996.tb01117.x

Jeong S (2016). How to make a bot that isn't racist: What Microsoft could have learned from veteran botmakers on Twitter. Vice. https://www.vice.com/en/article/mg7g3y/how-to-make-a-not-racist-bot. Accessed 27 Dec 2022

Johnson DG (2006) Computer systems: moral entities but not moral agents. Ethics Inf Technol 8(4):195–204. https://doi.org/10.1007/s10676-006-9111-5

Kahn PH, Kanda T, Ishiguro H, Gill BT, Ruckert JH, Shen S, Gary HE, Reichert AL, Freier NG, Severson RL (2012) Do people hold a humanoid robot morally accountable for the harm it causes? In: 2012 7th ACM/IEEE international conference on human-robot interaction (HRI 2012), pp 33–40. https://doi.org/10.1145/2157689.2157696

Kahn PH Jr, Kanda T, Ishiguro H, Gill BT, Shen S, Gary HE, Ruckert JH (2015) Will people keep the secret of a humanoid robot?—Psychological intimacy in HRI. In: 2015 10th ACM/IEEE international conference on human-robot interaction (HRI 2015), pp 173–180. https://doi.org/10.1145/2696454.2696486

Kim M (2017). Let robots handle your emotional burnout at work. How We Get to Next. https://www.howwegettonext.com/let-robots-handle-your-emotional-burnout-at-work/. Accessed 27 Dec 2022

Kim T, Song H (2021) How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. Telemat Inform 61:101595. https://doi.org/10.1016/j.tele.2021.101595

Lazare A (2004) On apology. Oxford University Press, Oxford

Lee MK, Kiesler S, Forlizzi J, Srinivasa S, Rybski P (2010) Gracefully mitigating breakdowns in robotic services. In: 2010 5th ACM/IEEE international conference on human-robot interaction (HRI 2010), pp 203–210. https://doi.org/10.1109/HRI.2010.5453195

Lee P (2016) Learning from Tay's introduction. Official Microsoft Blog. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/. Accessed 27 Dec 2022

Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183. https://doi.org/10.1007/s10676-004-3422-1

Morse B (2022) Sara Takanashi: Japanese ski jumper apologizes amid 'too big' suit disqualification controversy. CNN. https://edition.cnn.com/2022/02/09/sport/ski-jumping-women-disqualified-olympics-spt-intl/index.html. Accessed 27 Dec 2022

Noorman M (2020) Computing and moral responsibility. In: Zalta EN (ed) Stanf Encycl Philos (Spring 2020 edn). https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/. Accessed 27 Dec 2022

Okada M (2012) Yowai robot. Igaku Shoin, Tokyo (**Japanese**)

Pompe BL, Velner E, Truong KP (2022) The robot that showed remorse: repairing trust with a genuine apology. In: 2022 31st IEEE international symposium on robot and human interactive communication (RO-MAN 2022), pp 260–265. https://doi.org/10.1109/RO-MAN53752.2022.9900860

Reeves B, Nass C (1996) The media equation: how people treat computers, television and new media like real people and places. Cambridge University Press, Cambridge

Robinette P, Howard AM, Wagner AR (2015) Timing is key for robot trust repair. In: Tapus A, André E, Martin JC, Ferland F, Ammi M (eds) Social robotics: international conference on social robotics (ICSR) 2015. Springer, Cham, pp 574–583. https://doi.org/10.1007/978-3-319-25554-5_57

Rodogno R (2016) Social robots, fiction, and sentimentality. Ethics Inf Technol 18(4):257–268. https://doi.org/10.1007/s10676-015-9371-z

Scott NB, Lyman SM (1968) Accounts. Am Sociol Rev 33(1):46–62

Searle JR (1969) Speech acts. Cambridge University Press, Cambridge

Searle JR (1979) Expression and meaning: studies in the theory of speech acts. Cambridge University Press, Cambridge

Serholt S, Ljungblad S, Bhroin NN (2022) Introduction: special issue—critical robotics research. AI Soc 37(2):417–423. https://doi.org/10.1007/s00146-021-01224-x

Shank DB, DeSanti A, Maninger T (2019) When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. Inf Commun Soc 22(5):648–663. https://doi.org/10.1080/1369118X.2019.1568515

Sharkey N, Sharkey A (2006) Artificial intelligence and natural magic. Artif Intell Rev 25(1–2):9–19. https://doi.org/10.1007/s10462-007-9048-z

Smith N (2008) I was wrong: the meanings of apologies. Cambridge University Press, Cambridge

Sparrow R (2002) The march of the robot dogs. Ethics Inf Technol 4(4):305–318. https://doi.org/10.1023/A:1021386708994

Sparrow R (2007) Killer robots. J Appl Philos 24(1):62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sparrow R, Sparrow L (2006) In the hands of machines? The future of aged care. Minds Mach 16(2):141–161. https://doi.org/10.1007/s11023-006-9030-6

Strawson P (1962) Freedom and resentment. Proc Br Acad 48:187–211

Sweeney P (2022) Why indirect harms do not support social robot rights. Minds Mach 32:735–749. https://doi.org/10.1007/s11023-022-09593-y

Tavuchis N (1991) Mea culpa: a sociology of apology and reconciliation. Stanford University Press, Stanford

Turkle S (2011) Alone together: why we expect more from technology and less from each other. Basic Books, New York

Uchida T, Minato T, Koyama T, Ishiguro H (2019) Who is responsible for a dialogue breakdown? An error recovery strategy that promotes cooperative intentions from humans by mutual attribution of responsibility in human-robot dialogues. Front Robot AI 6:29. https://doi.org/10.3389/frobt.2019.00029

Verbeek PP (2011) Moralizing technology: understanding and designing the morality of things. University of Chicago Press, Chicago

Verbeek PP (2014) Some misunderstandings about the moral significance of technology. In: Kroes P, Verbeek PP (eds) The moral status of technical artefacts. Springer, Cham, pp 75–88. https://doi.org/10.1007/978-94-007-7914-3_5

Yamada K (2016) AI no idenshi. Vol. 5. Akita Shoten, Tokyo (**Japanese**)

Yamaji Y, Miyake T, Yoshiike Y, De Silva RS, Okada M (2011) STB: child-dependent sociable trash box. Int J Soc Robot 3(4):359–370. https://doi.org/10.1007/s12369-011-0114-y

Zeng V (2015, updated 2020) Robot of Japanese PM 'bows in apology to China' at Shanghai exhibition. Hong Kong Free Press. https://hongkongfp.com/2015/07/15/robot-of-japanese-pm-bows-in-apology-to-china-at-shanghai-exhibition/. Accessed 27 Dec 2022