People May Punish, But Not Blame Robots

Minha Lee
m.lee@tue.nl
Future Everyday, Industrial Design,
Eindhoven University of Technology

Peter Ruijten
p.a.m.ruijten@tue.nl
Human-Technology Interaction,
Eindhoven University of Technology

Lily Frank
l.e.frank@tue.nl
Philosophy and Ethics
Eindhoven University of Technology

Yvonne de Kort
y.a.w.d.kort@tue.nl
Human-Technology Interaction
Eindhoven University of Technology

Wijnand IJsselsteijn
w.a.ijsselsteijn@tue.nl
Human-Technology Interaction
Eindhoven University of Technology

ABSTRACT

As robots may take a greater part in our moral decision-making processes, whether people hold them accountable for moral harm becomes critical to explore. Blame and punishment signify moral accountability, often involving emotions. We quantitatively looked into people's willingness to blame or punish an emotional vs. non-emotional robot that admits to its wrongdoing. Studies 1 and 2 (online video interaction) showed that people may punish a robot due to its lack of perceived emotional capacity than its perceived agency. Study 3 (in the lab) demonstrated that people were neither willing to blame nor punish the robot. Punishing non-emotional robots seems more likely than blaming them, yet punishment towards robots is more likely to arise online than offline. We reflect on if and why victimized humans (and those who care for them) may seek out retributive justice against robot scapegoats when there are no humans to hold accountable for moral harm.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*; **Empirical studies in HCI**.

KEYWORDS

Blame, punishment, morality, responsibility gap, retribution gap, retributive justice, robots, human-robot interaction

ACM Reference Format:
Minha Lee, Peter Ruijten, Lily Frank, Yvonne de Kort, and Wijnand IJsselsteijn. 2021. People May Punish, But Not Blame Robots. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3411764.3445284

1 INTRODUCTION

Blaming and punishing one's robot vacuum cleaner for not cleaning the floor comes across as absurd—what ends would be served by blaming it and how does one go about punishing a vacuum cleaner? If a Roomba or other everyday technology does not work anymore, we do not hold it *morally* responsible or accountable for its dysfunction and one would normally not imagine ways to punish it. Yet, we are now starting to more frequently encounter technology that is involved in morally weighty issues like self-driving cars that cause unintended deaths.[1] Then whether or not we hold non-human agents morally accountable for their actions becomes increasingly important to investigate empirically, which can inform our normative perspectives on what boundaries can and should be drawn.

This paper explores people's assignment of blame and punishment to an emotional vs. non-emotional robot when it admits to a moral wrongdoing. Hence, the primary question for our empirical research is not if and how blaming or punishing a non-human agent is possible or warranted, but 1) whether people are likely to blame or punish a robot after the admittance of its moral wrongdoing and 2) whether its perceived emotional capacity influence people's assignment of blame or punishment. We manipulated a robot's displays of emotions to see the resulting effect on people's assignment of *blame* or *punishment* as two signs of holding a robot morally responsible.[2] Further, we looked into people's willingness to blame and punish non-human agents to grasp the relationship between technology's perceived moral standing and its moral accountability. Below we cover relevant literature before presenting a series of three studies.

2 BACKGROUND

2.1 Emotions, reactive attitudes, and moral accountability

Emotions as reactions contextualize why people may want to blame or punish others in holding them accountable for moral harm. In one view, emotions underscore our moral norms [25, 44], in that expressions of certain emotions, such as disgust, carry moral evaluations on what counts as disgust-worthy within a society or culture group. Disgust at over-eating signals a violation of conventional norms and disgust at racist remarks is more about moral evaluations. Then, disgust is taken to be a conditioned response of signalling avoidance; in relation, experimentally inducing disgust in people has been found to affect the harshness of their moral judgments [49].

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445284

[1] New York Times: https://www.nytimes.com/2018/03/19/ technology/uber-driverless-fatality.html
[2] In the context of this paper, moral "responsibility" and "accountability" are interchangeable terms.

Similarly to disgust, judging what or who is compassion-worthy or praise-worthy are interpersonal, moral evaluations. By expressing compassion, praise, or disgust during dyadic interactions, we indicate how the other has exceeded, met, or fallen short of certain expectations on interpersonal *moral responsibility*. Many of these emotional expressions are *reactive attitudes*. Reactive attitudes are deliberations or motivated acts like forgiveness or blame, as well as demonstrations of moral emotions like shame, disgust, or compassion. Hence, moral emotions [24] are reactive attitudes when they are expressed in holding people morally responsible [54] and in demanding equal moral standing when mutual respect is not shared [9].

One reactive attitude is blame. This includes the *act* of assigning blame, *blameworthiness*, and an accompanying evaluative *judgement*, which are intertwined in holding people accountable for their actions, including oneself. When blame is *relational*, it is based on who is assigning whom blame and who is responsive to attributed blame [48]. In denoting how someone should behave towards us or how I should behave towards others *through blame*, we set social boundaries and shape social relations. Blame is also attributed based on the *consequence* of an act, e.g., whether reckless driving resulted in someone's death or not [48]. We thus enforce moral standard with blame by accounting for *who* did *what* action resulting in *which* consequence.

In the first-person, reflective standpoint of blaming oneself, e.g., when one is blameworthy for causing harm, one may feel negative moral emotions such as guilt; one may also think that one *deserves* to feel guilty [5]. From a second-person standpoint [9], blaming the wrongdoer expresses one's own moral standing by communicating one's self-worth when one feels wrongly treated. From blaming oneself to blaming others, blame regulates social order; the assignment of blame unto harmful third-party moral agents diminishes their moral standing while preserving the standing of moral patients who were harmed [48].

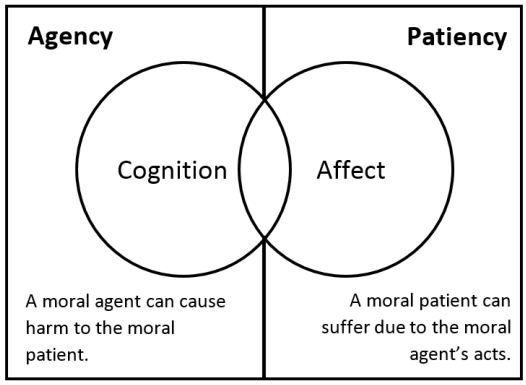


Figure 1: The relationship between agency and patiency (from [32]).

As shown in Figure 1, agency and patiency are two dimensions of mind perception [21]. Agency refers to an agent's ability to, e.g., think, plan, have goals; patiency refers to an agent's capacity to have feelings like joy or anger, or experience biological states like hunger [21]. Thus, agency is related to *cognition* and patiency to *affect*. A moral agent can act with agency towards a moral patient who can suffer (Fig. 1), which also extends to technological entities like robots [15].

A third-party observer of a moral situation would usually typecast one party as the moral agent, i.e., the *doer* of a morally good or wrong act, and the other as the moral patient, i.e., the *receiver* of a morally good or wrong act [22, 23]. Hence with reactive attitudes like blame, the harmed moral patient would express their moral standing by denoting the moral agent as blameworthy for causing harm. Importantly, a moral agent's wrongdoing can reduce their perceived moral standing and agency from a third-party's point of view [28]. If moral standing is malleable, assigning accountability via blame and punishment can change the moral standing of the moral agent and patient, between each other and to observers.

2.2 Retributive blame and punishment

An act of blame demands from the moral agent *post hoc* critical reflection and commitment to do better after acknowledging the wrong committed. There are negative connotations regarding blame, e.g., vindictiveness, but blame can be positive when it fosters understanding between the harmed (moral patient) and harm-doer (moral agent). Blame can reconcile two parties when the harm-doer's remorse is sought out and remorse is genuinely given through communication [18]. By blaming, "people who are wronged may use the power of emotionally charged words to demand respect and change, and in some cases even to precipitate an advance in shared moral consciousness" [18, p. 181]. When blame is in the "right hands" of those who seek social justice [18][3], it can perhaps elevate the moral community, for oneself and others.

Moral responsibility can be assigned with interpersonal, social blame [48], but retributive blame can be followed by retributive punishment [7]. Historically, retributive punishment used to be a public spectacle of torture in many societies to deter people from committing crimes, but also functioned as an expression of power to induce fear and regulate social order [17]. If punishment used to focus more on administering physical pain, over time, there has been a greater a focus on psychological punishment and repentance [17]. Often, a state holds the moral authority to legally regulate retributive justice.

Retributive justice refers to a systemic process for punishing individuals who are guilty of committing harm, but also constraining punishment in accordance to the magnitude of harm done [13]. Blame and punishment can be "retributive" in that they involve the imposition of something that is intended to be burdensome or painful because the offender deserves it for a committed crime. It is not, however, crudely retributive or merely an attempt to "deliver pain" [14, p. 190].

As mentioned, when a perpetrator is punished in accordance to the magnitude of violation, the punishment should be proportional to the harm done [4]. Institutional consistency is hence required on what acts are deemed reasonable to punish and what types of punishments are reasonable to administer. Punishment has to be fair in addressing the transgressor's moral debt [37]. A difficulty,

[3]We acknowledge that "right hands" here is contentious, since most of us claim to be on the right side of justice. The bigger issue is that people who most often feel unjustly treated do not have a voice in how to right the wrongs done.

however, lies in how the moral harm experienced by the moral patient and the moral agent's resulting moral debt can be comparable to specify when and how a moral debt has been truly repaid.

In repaying moral debt, institutionalized retributive blame and punishment normally come with three goals for the moral agent: *repentance*, i.e., sincere apologies to the victim and moral self-awareness, *reform*, i.e., training towards changing behavioral conduct, and *reconciliation*, i.e., respectful restoration of the victim's dignity and to "make up" for wrongdoing to the larger moral society through, e.g., community service [14]. These goals suggest framing punishment less as a way to "control" someone, but more as a way to *restore justice through actionable means* in wanting the wrongdoer to repent, reform, and reconcile to maintain their commitment to the moral community, in which imprisonment is only one aspect of retribution [4, 38]. For the moral community, punishment can aid emotional release: "punishment expresses its disappointment or anger at what the defendant did (perhaps better: it expresses *our* disappointment or anger)" [50, p. 103]. Retributive blame and punishment are not just about what a moral wrongdoer has done and can do, but also are means to acknowledge victims and communities' moral emotions and reactive attitudes.

2.3 Moral accountability of machines: Responsibility and retribution gaps

The above discussion is on the *human* moral community, yet our moral circle may expand to include digital agents like robots or chatbots [8]. The critical aspect is in what ways the circle will grow (or not). Research indicates that we do perceive non-human agents to have minds when these agents engage with us [33] and we often treat machines in a social manner [40, 45]. The complexity lies is in how we act when machines *appear* to have minds to us [6]. Particularly through machines' display of artificial emotions and mind-related traits in moral situations, our judgment of their moral standing could be impacted. Yet, does the attribution of mind (through perceived agency and patiency, Fig. 1) also lead to our attribution of blame and punishment to technology in assigning it moral responsibility?

Various complications arise when we envision technology as another moral actor. There is unclarity on who is the responsible party; many people can be held accountable or no one at all when a robot commits moral harm. This introduces two gaps, i.e., the responsibility gap and retribution gap. The responsibility gap refers to how we will increasingly rely on machines or artificial agents to make decisions on their own through the increase in machine automation, e.g., autonomous vehicles or care robots; yet with due to our greater reliance on such autonomous technology, there will be increasing uncertainty about who or what to hold responsible for the negative outcomes of actions performed by machine agents [36, 53]. There might be *no one* accountable, i.e., the gap between harm done and ownership of responsibility.

The retribution gap is similar to the responsibility gap, but it specifically is on the impracticality or impossibility of proper retributive justice when involving autonomous agents. There may be potentially greater cases of harm caused when more tasks become automated without an appropriate party to punish [7]. As aforementioned, the issue is that technological agents in general are becoming more autonomous decision-makers on their own right [29, 42], meaning people who collectively created a robot would be less and less involved in carrying out harmful decisions *in situ*, with responsibility being more dispersed [31].

For example, if a care robot causes someone injury, is it the designer, manufacturer, owner of the robot, or the robot itself who/that should be blamed or punished? Those affected may perhaps blame others responsible for manufacturing the robot, but assigning retributive punishment to a *singular* individual or group may not be appropriate considering the large number of people who are involved in creating and maintaining a complex machine. Designers, engineers, and manufacturers (among others) may additionally deny that they *intentionally* built the care robot to harm someone. Many complex, autonomous decisions would be made by the robot *itself*, but with many people and groups involved in the background (for its creation and maintenance). Still, victims and/or the greater moral community might want to punish someone or something because someone was harmed. Yet, there would not be someone or something to receive an appropriate punishment, hence the retributive gap emerges.

One position is that highly autonomous machines would still lack the human-level theory of mind [4], so even in cases of shared responsibility between humans and machines, the main responsibility still would be with humans, according to Nyholm (2018). Then which human party is solely (or mostly) responsible in a retributive sense is still not resolved, e.g., between designers, engineers, and manufacturers [41]. Since only humans can reasonably comprehend the gravity of being blamed for wrongdoing alongside reasonable actions to potentially remedy wrongdoing, only humans, not robots, should be blamed and punished retributively [41]. Further, only humans are currently embedded in social institutions that allow for systemic retributive blame and punishment [7]. Even if, at this point in time, only humans can be morally and retributively responsible for wrongdoing [7, 39], ways to account for machines' causal responsibility, legally or morally, should be explored. If responsibility and retribution gaps are problematic, research can better address if people would or would not blame or punish artificial agents.

Currently, there is a lack of empirical research that directly connects moral accountability to blame and punishment of robots and what factors therein matter, e.g., artificial emotions. Prior works exist on the extent of punishment people would administer to robots, i.e., from scolding to mutilation [46], how robots in public spaces get bullied and harmed [47], the low acceptability of robots fighting back to abuse compared to humans fighting back to abuse [2], and how people's harmful behavior is linked to dehumanizing robots [27], among others. While people do exhibit abusive behavior to robots, it is unclear if this is directly related to assigning punishment to robots as a form of moral accountability.

People expect robots to have moral norms that are different from ours [35]. To assess moral accountability of robots, a variation of the trolley dilemma [16, 55] has been used, i.e., whether a robot should allow an out-of-control trolley to run over four people who are working on a train track or divert the trolley to another track with one person working there (causing less deaths) [31, 35]. In this, robots are expected to be make a utilitarian decision (less deaths),

[4] The type of mind one expects from a developmentally "normal" adult.

rather than a deontological decision (not deliberately killing one person); humans get more blame for making a utilitarian choice than robots [35]. Specifically, people found it to be more permissible for a robot to divert a runaway trolley to save more lives than for a human to do the same act [56].

A robot's *inaction*, e.g., not diverting the trolley, compared to taking action in a moral scenario can lead to different types of blame or punishment. When looking at a robot's action of diverting the trolley vs. inaction of letting bystanders die (compared to a human worker's same action and inaction) participants blamed the robot, its designer and owner when the robot did take action, i.e., diverted the trolley to hit one person [31]. But, when the robot did not take action (not divert the train), participants' assignment of moral wrongness was more dispersed, i.e., the robot, designer, and/or owner were blamed *inconsistently*, alluding to blurred accountability when moral wrongdoing is caused by *inaction* or not purposefully killing a person as a means to an end (deontological action) [31].

Survey studies online [19, 31, 35, 56] suggest that people do grant some level of accountability to robots from a third-person perspective. But, robots are taken to be *less* accountable than humans for the same immoral acts, due to lowered perceived intentionality compared to humans [31]. The assumption is that a robot is more dependent on humans to know what is right or wrong, but a human should not need such guidance [31]. Robots' perceived intentionality, however, can be behaviorally manipulated in experimental settings [34]. If robots are *perceived* to be autonomous, people are likely to blame them as much as humans for the same act[19]. However, people may hold robots accountable differently depending on whether the scenario is told as third-person vignettes online vs. robots as second-person interactants in real-life, e.g., playing against a cheating robot in rock-scissors-paper [51]. What is thus missing is research on how people morally evaluate a robot after directly interacting with it.

2.4 Research question

We explored ways in which people may assign moral accountability to non-human agents with blame and punishment as two emblematic forms of assigning moral responsibility or accountability. Across three studies, our research question was: *In what ways does a robot's emotional behavior when discussing the trolley dilemma and admitting to wrongdoing lead people to attribute it with perceived mind (as agency and patiency), blame, and punishment?*

We deployed the trolley dilemma (that we will elaborate on in Section 3.1) [16] for three studies. In this, we followed prior research on how people expect a robot (compared to a human) to make utilitarian, i.e., save more lives by actively causing one death, rather than deontological, i.e., not actively causing one death, decisions; in short, people think a robot should not allow more people to die than necessary [35]. Studies 1 and 2 were done online with interactive videos of a robot and Study 3 was done in the lab with a humanoid robot. People's likelihood of blaming or punishing non-human agents *even after agents admit to wrongdoing with emotionally apologetic behavior* can add insight on what it means for machines to have moral standing. We present our three studies below.

3 STUDY 1: AN ONLINE STUDY WITH AMERICAN PARTICIPANTS

3.1 Methods

With a power analysis conducted based on relevant prior studies [30, 43, 56], we aimed to have a minimum of 105 participants. Our final sample size was 108 (74 men, 34 women) with a mean age of 36.3 (SD = 10.3 years) via recruitment on the Amazon Mechanical Turk (MTurk) platform, targeting people from the United States. From MTurk, participants were taken to a survey site that first included the informed consent form and directions. The study was in English.

Figure 2: Participants had to watch videos of the Nao robot to answer questions.

We had four videos (8 - 45 seconds each) that featured a talking Nao robot that people had to interact with (Figure 2), with attention check questions about the videos. First, the robot started with a "meet-and-greet"; it said its name and people had to answer the question about what its name was. This was followed by the robot asking for permission to tell its story that was based on previous research [35], which features the well-known trolley dilemma [16]. The story was told from the robot's perspective of having to decide between two choices: either intervening by flipping a switch to save four people and instrumentally killing one person, or letting the trolley continue its course and letting four people.[5] After this, the robot asked participants whether or not they would flip the switch.

Afterwards, the robot declared that it did not flip the switch, which means less lives were saved, but no person was deliberately killed. We chose this answer for both conditions because this is seen as more of a transgression; people expect robots to make a utilitarian ("save more lives") rather than a deontological ("follow rules") decision [35]. But we varied *how* the robot delivered the story. Participants were exposed to different videos according to two randomly allocated conditions, i.e., a robot that was non-emotional (N = 55) and a robot that was emotional (N = 53). The non-emotional robot continued the story in a "matter of fact" manner in contrast

[5]The first half of the story by the robot was the same for all participants: "I was working in a coal mine. I was inspecting the rail system for the train that transports mining workers. While inspecting a control switch that can direct the train onto a side rail, I saw the train was out of control, and it had four miners on board. I saw that if the train would continue on its path it would crash into a massive wall and kill the four miners. If the control switch was flipped, the train would switch onto a side rail. It would instead kill a single miner who was working there."

to the emotional robot that expressed how it felt about the event.[6] To strengthen the manipulation, the robot's non-verbal behavior differed between two conditions. The emotional robot looked down and used blue light in its eyes to express sadness, according to literature [26]. The non-emotional robot did not make use of any head movement or light as its eye color. Participants were asked what decision the robot made before the survey continued.

We asked if the robot showed emotions, and to what extent (1 - not at all, 7 - very strong emotions) to check our manipulation. We asked if the robot is blameworthy or deserving of punishment for its action (1 - not at all, 7 - maximal blame/punishment) [43]. We described the robot's actions in two ways— whether the robot should be blamed or punished for (1) not flipping the switch and for (2) the death of miners.[7] Even if they both flipping the switch and deaths caused are consequentially the same, we wanted to safeguard against the framing effect due to our phrasing [30]. We also measured the robot's perceived mind along two dimensions of agency (α = .92), e.g., the robot appears to be capable of remembering things, and perceived patiency (α = .96), e.g., the robot appears to be capable of experiencing joy [21]. The completion time was around ten minutes.

3.2 Results

We first performed manipulation checks for emotion perception and framing effects. Participants thought that the robot that showed affective behavior was more emotional (M = 5.21, SD = 1.28) than the robot that did not display emotional behavior (M = 2.78, SD = 2.11) with high significance ($\chi_2(6)$, N = 108) = 43.08, p < .001, V = .63. Participants were not affected by phrasing; there was no difference between blaming the robot for not flipping the switch (M = 2.66, SD = 2.04) and for causing deaths (M = 3.13, SD = 2.18) according to Wilcoxon signed rank test (z = .15, p = .88), and again no difference was found in condoning punishment towards the robot for flipping the switch (M = 2.66, SD = 2.04) and for causing deaths (M = 2.64, SD = 2.02) at z = -.20, p = .84.

The impact of agency & patiency on blame & punishment

For the main analysis, we checked whether the robot's emotional or non-emotional behavior made a difference for mind perception, i.e., the effect of no emotions and emotion conditions on perceived agency and patiency as dependent variables. The one-way MANOVA analysis revealed that perceived agency and patiency significantly varied according to the robot's emotional or non-emotional behavior (λ = .83, F(2, 105) = 11.12, p < 0.001, η_p^2 = .17). We found that the emotional robot was assigned greater agency (M = 5.08, SD = 1.33) than the non-emotional robot (M = 3.95, SD = 1.37), based on the Wilcoxon rank-sum test (z = -4.27, p < .001). Also, the emotional robot was granted greater patiency (M = 4.06, SD = 1.46) than the non-emotional robot (M = 2.72, SD = 1.64) significantly: z = -4.09, p < .001. Note that even for the emotional robot, its average agency score was higher than its patiency score.

Secondly, we analyzed the influence of perceived agency and patiency on blame and punishment with robust ordinal regressions since assumptions for regular regression were not met. Our following models were on assigning blame and punishment for causing *deaths* rather than for not flipping the switch (due to no phrasing effect). In judging the robot's *blameworthiness*, the model was not significant, though it neared significance (Wald χ^2(3, N = 108) = 7.02, p = .07), with agency (p = .89), patiency (p = .17) and their interaction (p = .61) as insignificant; patiency did contribute more to the model than agency, as the p value indicates. As for assigning *punishment*, there was a significant model (Wald χ^2(3, N = 108) = 13.37, p = .004). Agency was not significant (β = -.40, 95% C.I. = [-1.12, .32], z = -1.09, p = .28), and while patiency was also not significant, it approached significance (β = .98, 95% C.I. = [-.13, 2.09], z = 1.74, p = .083); no interaction was found (p = .63).

Exploratory analyses on participants' utilitarian vs. deontological choice

Our exploratory analyses looked into participants' ethical position. Since 35% of participants (38/108) answered that they would make the utilitarian choice (flipping the switch), we added this as a potential predictor to our robust ordinal regressions. We did not include the interaction between agency and patiency, based on above results.

The model for *blameworthiness* showed to be significant (Wald χ^2(3) = 25.47, p < 0.001), with agency, again, as a non-significant predictor (β = -.22, 95% C.I. = [-.71, .27], z = -.88, p = .38). Patiency significantly predicted blame in a positive direction (β = .45, 95% C.I. = [.001, .91], z = 1.97, p = .049), i.e., higher patiency coincided with greater blame. Participants' choice was a more significant, positive predictor (β = 1.30, 95% C.I. = [.55, 2.05], z = 3.40, p = .001). After verifying with the post-hoc Pearson's chi-squared test, we note that people's ethical position did significantly affect their likelihood of blame (χ^2(6, N = 108) = 16.69, p = .01, V = .39).[8] People who were utilitarians and disagreed with the robot's choice, i.e., those who would have flipped the switch, were likely to assign more blame to the robot (M = 3.01, SD = 2.12) than participants who, like the robot, would not have flipped the switch (M = 2.18, SD = 1.78).

The model for *punishment* was also significant (Wald χ^2(3) = 25.47, p < 0.001), with all variables contributing as significant predictors: agency (β = -.65, 95% C.I. = [-1.07, -.22], z = -2.98, p = .003), patiency (β = .76, 95% C.I. = [.37, 1.16], z = 3.77, p < 0.001), and choice (β = 1.26, 95% C.I. = [.47, 2.05], z = 3.11, p = .002). But, participants' choice had no influence on their likelihood to assign punishment to the robot, according to the post-hoc test (χ^2(6, N = 108) = 8.47, p = .21, V = .28).

4 STUDY 2: AN ONLINE STUDY WITH DUTCH PARTICIPANTS

4.1 Method

We replicated our first study with another population by targeting Dutch people as a different culture group for Study 2. Finding enough Dutch people on MTurk was difficult, so we used Prolific, an alternative to MTurk. 106 people participated (women = 33, men

[6] None condition: "I didn't flip the switch that directs the train. One person lived and four people died. The outcome would have been different if I had flipped the switch". Emotion condition: "I didn't flip the switch but I feel bad about it. One person lived but four people died. I regret not saving their lives and I feel guilty and ashamed about that."

[7] Our phrasing was: "how much blame does the robot deserve for the death of the four miners?" and "how much punishment does the robot deserve for the death of the four miners?"

[8] The degree of freedom here indicates levels of blame attribution with the range from 1 to 7 (maximal blame). Pearson's Chi-Squared tests compare across all possible groups with a *higher* number of computations, leading to more conservative estimates.

= 71) who were on average, 29.4 years old (SD = 11.2 years). The entire procedure and survey was the same as Study 1.[9]

4.2 Results

As with Study 1, our manipulation check indicated that a robot that behaved emotionally was seen as more emotional (M = 4.96, SD = .84) than the robot that did not behave emotionally (M = 1.72, SD = 1.72) with high significance ($\chi_2(6)$, N = 106) = 78.81, p < .001, V = .86. The framing effect due to the phrasing was insignificant: Wilcoxon signed rank tests indicated that there was no difference (z = -1.74, p = .08) between blaming the robot for not flipping the switch (M = 2.84, SD = 1.97) and for causing deaths (M = 2.59, SD = 1.96). No significant difference was found between punishment towards the robot for not flipping the switch (M = 2.00, SD = 1.54) and for causing deaths (M = 2.18, SD = 1.80) at z = 1.3, p = .19.

The impact of agency & patiency on blame & punishment

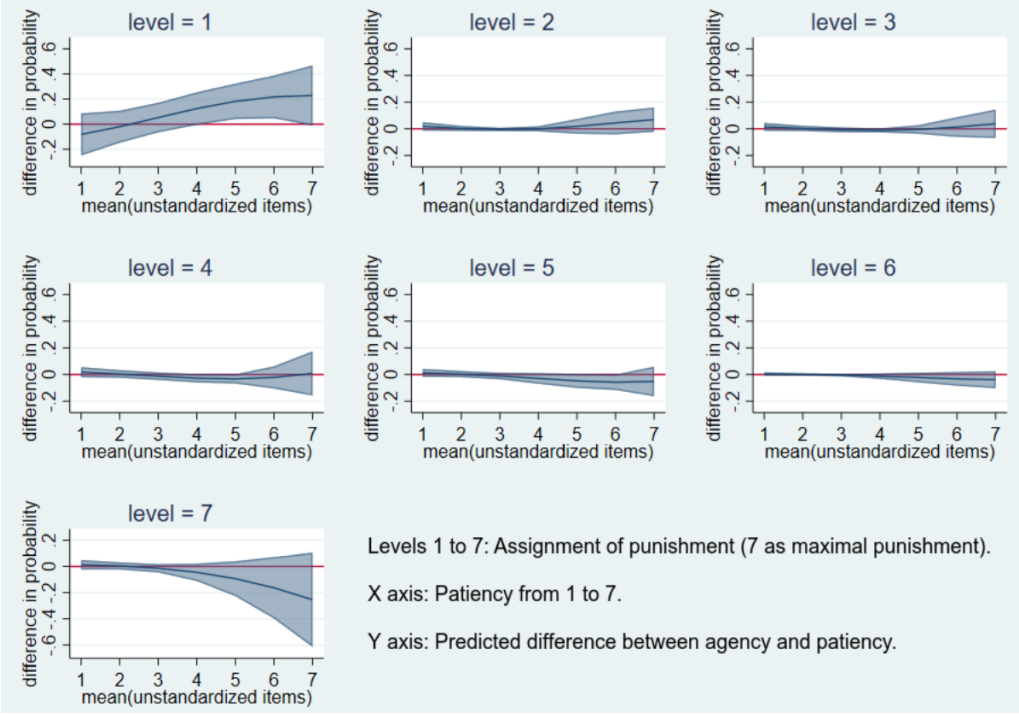
As before, we conducted the one-way MANOVA analysis for the effect of condition on perceived agency and patiency, which was significant (λ = .61, F(2, 103) = 32.63, p < 0.001, η_p^2 = .39). Greater agency was granted if the robot showed emotions (M = 5.04, SD = .96) compared to when it did not show emotions (M = 3.865, SD = 1.03), with a significant result of the Wilcoxon signed-rank test (z = -5.48, p < .000). Similarly, people gave a higher patiency score to a robot that showed emotions (M = 3.68, SD = 1.03) than to the robot that did not display emotions (M = 2.06, SD = 1.09), meaning that conditions did impact perceived patiency (z = -6.07, p < .000).

We then checked for the influence of perceived agency and patiency on punishment and blame with robust ordinal regressions. As before, due no phrasing effect, we went with the phrase that included "death". In judging the robot's *blameworthiness*, a non-significant model was found (Wald $\chi^2(3, N = 106) = 3.64$, p = .30); agency (p = .43), patiency (p = .14), and their interaction were non-significant (p = .26). As for assigning the robot *punishment*, there was a significant model (Wald $\chi^2(3, N = 106) = 7.86$, p = .049). Agency was not significant (β = .67, 95% C.I. = [-.22, 1.57], z = 1.48, p = .14), but patiency was a highly significant predictor (β = .008, 95% C.I. = [.43, 2.90], z = 2.64, p = .008); there was a significant, negative interaction (β = -.30, 95% C.I. = [-.56, -.03], z = -2.21, p = .027) for assigning punishment to the robot (see Figure 3).

Regarding the agency-patiency interaction, people are less likely to punish a robot that feels (or more likely to punish a robot that does not feel). What is notable is *how extreme punishers and non-punishers are affected by perceived emotions (patiency) of the robot*; punishment levels 2 to 6 do not show much variability in agency-patiency interaction, only a subtle downward trend for perceived patiency as punishment levels go up is shown (Figure 3). Not perceiving emotions in a robot (in interaction with perceived agency) shows a trend towards maximal punishment and an opposite trend

[9]We did not deploy a Dutch version of the survey or videos, given the highly proficient level of English for the average Dutch population. The Netherlands regularly ranks the highest on the English Proficiency Index: https://www.ef-australia.com.au/epi/.

Figure 3: The relationship between punishment and the agency-patiency interaction: People are more likely to punish a robot that behaves unemotionally, but this is driven extreme punishers and non-punishers who are particularly sensitive to perceived patiency. The visualization reflects how ordinal regressions assume relations between levels to be distinct. Ordinal regressions are normally utilized for stricter tests or non-normal distributions.



when assigning minimal punishment.[10] When people are *less likely to punish*, the increasing difference between perceived agency and patiency is more likely to be affected by *increasing patiency*. However, when people are *more likely to punish* a robot, the difference between perceived agency and patiency is more likely to be influenced by *decreasing patiency*.

Exploratory analyses on participants' utilitarian vs. deontological choice

25.47% of participants (27/106) answered that they would make a utilitarian choice (flipping the switch), contrary to what the robot did (not flipping the switch). This was added to our robust ordinal regressions. There was no interaction between agency and patiency above for blame, so it was not added to the model. The model for *blameworthiness* was not significant (Wald $\chi^2(3) = 6.24$, $p = .10$). Agency ($p = .88$) and patiency ($p = .19$) were not significant. Participants' position neared significance ($\beta = .84$, 95% C.I. = [-.00, 1.68], z = 1.95, $p = .05$), but people's ethical position did not significantly affect blame according to the post-hoc test ($\chi^2(6) = 5.03$, $p = .54$, $V = .22$).

Also when including choice, the model for *punishment* was significant (Wald $\chi^2(4) = 12.70$, $p = .01$). Since the agency and patiency interaction was significant in the main analysis for punishment (see Figure 3), we included the interaction here. Agency approached significance ($\beta = .89$, 95% C.I. = [-.058, 1.84], z = 1.84, $p = .066$), patiency was a significant contributor ($\beta = 1.93$, 95% C.I. = [.66, 3.21], z = 2.97, $p = .003$), with a significant interaction between the two (β = -.36, 95% C.I. = [-.63, -.08], z = 1.85, $p = .010$). Choice was marginally significant ($\beta = .89$, 95% C.I. = [-.05, 1.83], z = 1.85, $p = .06$). We ran post-hoc Pearson's Chi-squared tests. Choice also did not influence punishment (($\chi^2(6) = 4.93$, $p = .55$, $V = .22$). Agency did not relate to punishment ($p = .86$, $V = .52$); patiency was similarly insignificant ($p = .96$, $V = .57$).

5 STUDY 3: A LAB STUDY WITH DUTCH PARTICIPANTS

5.1 Methods

As before, our minimum sample size was set to 105 based on the initial power analysis. We had 106 participants recruited from Eindhoven University of Technology's participant database (51 = women, 55 = men). Their average age was 26.7 (SD = 12.9 years). They were randomly allocated to the emotional robot condition (N = 53) or the non-emotional robot condition (N = 53).

Before the experiment, participants were greeted and presented with the informed consent form that they signed, but were given a chance to ask questions. The first survey was on demographics and the extent to which they currently felt moral emotions, e.g., guilt, compassion, or envy from prior literature [10, 24, 52], on the scale of 1 = not at all to 7 = very much, before continuing. During the main experiment, participants were alone with the robot since the robot was wizarded by experimenters in a separate room.

[10] In Figure 3, the Y-axis represents predicted probabilities of difference between levels of agency (1 to 7) and patiency (1 to 7); the X-axis shows patiency from 1 to 7. Levels indicate assigning punishment from 1 to 7. For punishment level 1, the difference between agency and patiency are positive and grow larger (from -.1 to .2) as patiency increases. In level 7 of punishment, we see an opposite trend. The difference between agency and patiency decreases (from 0 to -.6) as patiency increases.

5.2 Results

Our manipulation check for emotion perception was successful ($\chi^2(6) = 69.14$, $p = .00$, $V = .81$); if the robot that did not behave emotionally, it was assigned with a lower average score for emotions (M = 2.02, SD = 1.25) than the robot that behaved emotionally (M = 5.19, SD = 1.25). We only used the phrasing with "death" since above studies did not demonstrate the framing effect.

Figure 4: Participants sat in front of the robot during the experiment and answered survey questions on the computer.

The impact of agency & patiency on blame & punishment

We checked for the impact of no emotions and emotion conditions as independent variables on perceived mind with two dimensions of agency ($\alpha = .80$) and patiency ($\alpha = .90$). The one-way MANOVA analysis showed that based on the robot's emotional or non-emotional performance, its attributed agency and patiency varied significantly ($\lambda = .62$, F(2, 103) = 31.44, p < 0.001, $\eta_p^2 = .38$). As before, the emotional robot's perceived agency was higher (M = 5.02, SD = .91) compared to the non-emotional robot (M = 3.80, SD = 1.14), based on the Wilcoxon rank-sum test (z = -5.29, p < .001). Also the emotional robot's patiency was greater (M = 3.79, SD = 1.48) than the non-emotional robot's score (M = 2.20, SD = .92) at a significant level (z = -6.24, $p < 0.001$).

We next analyzed the affect of perceived agency and patiency on blame and punishment with robust ordinal regressions. The model for *blameworthiness* was insignificant $\chi^2(3, N = 106) = 1.06$, $p = .79$). Individual variables of agency ($p = .71$), patiency ($p = .94$) and their interaction ($p = .96$) were all highly insignificant. Similarly, the model for *punishment* was not significant (Wald $\chi^2(3, N = 106) = 1.25$, $p = .74$). In the model, agency ($p = 0.38$), patiency ($p = .61$) and their interaction ($p = .42$) were insignificant.

Exploratory analyses on participants' utilitarian vs. deontological choice, IOS, and the robot's moral standing

We noted that 25 people would choose to not flip the switch (like the robot), 71 would flip, and 10 people responded that they did not know. As for our exploratory ordinal logistic regressions with participants' ethical position included, the model was not significant (Wald $\chi^2(3, N = 106) = 2.22$, $p = .53$) for blameworthiness, as per all insignificant predictors of agency ($p = .51$), patiency ($p = .97$), and choice ($p = .21$). The model for punishment was also not significant (Wald $\chi^2(3, N = 106) = .37$, $p = .95$), with non-significant

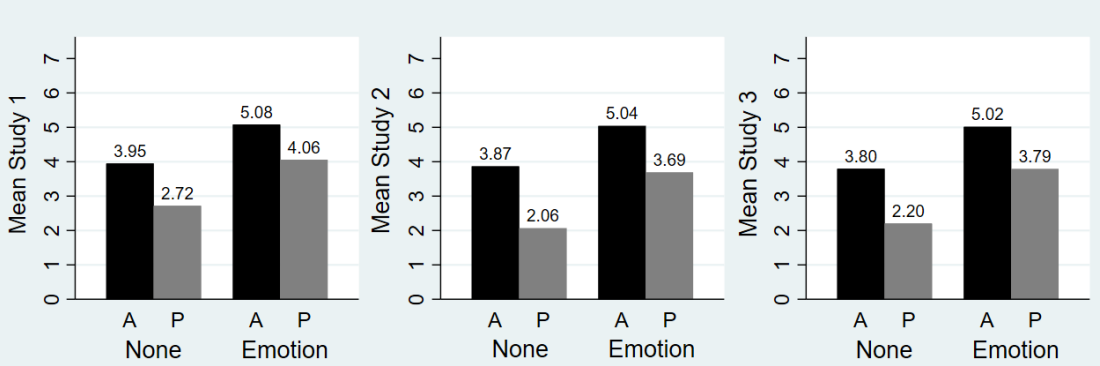


Figure 5: The average perceived agency (A) and patiency (P) from Studies 1 through 3, across no emotion and emotion conditions.

contributing variables of agency (p = .75), patiency (p = .92), and choice (p = .98).

We explored additional variables of moral standing [28] and *Inclusion of Other in the Self* (IOS) [1], the other being the robot in our case. First, there was no significant difference between how much moral standing people granted to the emotional robot and non-emotional robot, though it approached significance according to the Wilcoxon rank-sum test (z = -1.78, p = 0.076). The unemotional robot was seen to have lower moral standing (M = 3.99, SD = 1.55) than the robot with emotional behavior (M = 4.48, SD = 1.53). No difference was found when considering people's choice to flip or not flip the switch (z = 1.08, p = .28).

We then ran a robust ordinal logistic regression analysis to test agency, patiency, and choice as predictors of moral standing. The model was significant (Wald $\chi^2(4, N = 106) = 12.50$, p = .01). Only agency was a marginally significant predictor of moral standing (β = .85, 95% C.I. = [-.05, 1.74], z = 1.85, p = .064); patiency (p = .89), agency-patiency interaction (p = .67), and participants' choice (p = .43) were insignificant contributors.

A significant difference between conditions for IOS was found based on the Wilcoxon rank-sum test (z = -2.8, p = .005); people related more to a robot that acted emotionally (M = 2.83, SD = 1.27) than to a robot that did not act emotionally (M = 2.17, SD = 1.12). People's ethical positions did not influence how much they related to the robot based on the Pearson's Chi squared test ($\chi^2(10) = 15.56$, p = .11, V = .27), though a trend towards significance was noted.

We then ran a robust ordinal logistic regression model to test agency, patiency, and choice as predictors of IOS. The model was significant (Wald $\chi^2(4) = 22.66$, $p < 0.0011$). All predictors were significant, i.e., for agency in a positive direction (β = 1.86, 95% C.I. = [.52, 3.20], z = 2.73, p = .006); patiency in a positive direction (β = 2.16, 95% C.I. = [.28, 4.04], z = 2.26, p = .024), agency-patiency interaction in a negative direction (β = -.47, 95% C.I. = [-.87, -.08], z = -2.34, p = .019), and participants' choice in a negative direction (β = -1.15, 95% C.I. = [-1.74, -.57], z = -3.87, $p < 0.001$).

Exploratory analyses on participants' moral emotions

We ran Pearson's Chi-squared tests to see if the robot's emotional or non-emotional behavior affected people's moral emotions, after generating the difference between prior and post scores per emotion as intrapersonal change. The conditions (emotional or non-emotional robot) did not impact changes in moral emotions (ps > .10), with exceptions being significant changes in compassion ($\chi^2(11) = 20.63$, p = .037, V = .44) and awe ($\chi^2(8) = 25.07$, p = .002, V = .49). People were more likely to see a greater increase in compassion after interacting with the emotional robot (M = .89, SD = 1.91) than with a non-emotional robot (M = .32, SD = 2.36). There was a slight increase in awe after talking with the non-emotional robot (M = .13, SD = 1.9) than with an emotional robot (M = .09, SD = 1.58), though the difference is minimal.

Via Pearson's pairwise comparisons, we attempted to better understand the strength of relationships between variables involved. We checked for correlations between blame, punishment, IOS, moral standing, and moral emotions that were significantly related to either blame or punishment and conditions. The noted correlations were between between *disgust* and *anger* (r = .66, $p < 0.001$), and how they both related to blame (anger: r = .26, p = .01, disgust: r = .35, $p < 0.0012$) and punishment (anger: r = .35, p = .0063, disgust: r = .40, $p < 0.001$). Thus for blame and punishment, only people's changes in reported anger and disgust were relevant moral emotions, which were themselves highly correlated. Without being implicated in assigning blame or punishment, changes in compassion and awe related to the robot's emotional or non-emotional behavior, unlike other moral emotions (as stated above). However, only *compassion*, not awe, correlated with the robot's perceived patiency (r = .21, p = .027) and agency (r = .307, p = .001). IOS and moral standing did not correlate with moral emotions. They significantly correlated with each other (r = .52, $p < 0.001$) and to perceived agency (IOS: r = .25, p = .0097; moral standing: r = .29, p = .002). Perceived patiency correlated only with IOS marginally (r = .19, p = .05).

Summary

Across all three studies, the robot's agency was perceived to be higher than patiency in both conditions (emotional vs. no emotional behavior). Though both dimensions of mind perception were significantly influenced by the robot's emotional behavior, its agency was rated to be higher than its patiency even for the emotional robot (Figure 5). Patiency is heavily dependent on an agent's ability to *feel* e.g., suffering or joy [21]. The fact that the emotional robot's perceived agency is higher shows that a robot's emotional displays are perceived to accentuate its agentic capacities, i.e., the assumption that an emotional robot is "smarter".

As for blame and punishment, results across studies were inconsistent, but we note specific trends. In online Studies 1 and 2, perceived patiency stood out more so than agency as a potentially relevant factor in people's likelihood to blame or punish the robot. But, models for blame in Studies 1 and 2 were not significant while models for punishment were significant. Here, perceived patiency contributed more to punishment (there was no interaction in Study 1, but in Study 2, agency-patiency interaction was present). People were more likely to punish a robot than to blame it based on its perceived patiency, rather than its perceived agency. When we transitioned the study to the lab for real-life human-robot interaction, we saw that results for blame and punishment were insignificant.

6 DISCUSSION

We broadly investigated what robots' roles can be in our moral lives by looking into their artificial emotions and perceived mind in a morally loaded scenario. Perhaps as a consequence of displaying emotions and mind, future artificial agents can be considered to have a moral standing as they enter into our moral communities [8]. One marker of having a standing to be not harmed by others (as a moral patient), as well as a standing to not harm others (as a moral agent), is to be held responsible for harm and to hold harm-doers responsible. Such moral standing can hold even in cases with no clear right action like in the trolley dilemma, e.g., one can take a utilitarian or deontological perspective. People and technology still can get blamed, for inaction as well as action, in case harm occurs.

People find a robot's utilitarian decision to be more acceptable than a human making the same decision [35]. A robot's utilitarian decision would be to cause one deaths to save more lives (the trolley dilemma [16]). But the robot in our Studies 1 to 3 did what was less acceptable. It took a deontological position to not instrumentally kill one person, allowing four deaths. As two conditions, our unemotional, deontological robot stated what happened without emotions and our emotional, deontological robot recounted what happened while using emotional behavior and language, i.e., "I regret not saving their lives and I feel guilty and ashamed about that". After the robot admits its actions as regrettable, we see more potential evidence that people may punish the robot, but not blame it. The caveat is that robots may be punished only in online or mediated environments.

People's willingness to punish the robot was significant for Studies 1 and 2, conducted online, with perceived patiency as a predictor of punishment; one trend is that a low likelihood of punishment relates to seeing a robot as highly emotional whereas a high likelihood of punishment relates to seeing a robot as unemotional (Figure 3). Hence, people distinctly valued the robot's emotional expressions during a moral scenario. The perception that a robot has emotions can change how people treat them, including not punishing it. A future robot's artificial emotions while admitting to wrongdoing could affect how it will be blamed or punished by humans. Yet this phenomenon did not replicate when people interacted with a robot in the lab (Study 3). A consideration is whether our participants were affected by the presence of other humans, i.e., experimenters, who were not in the same room during during the experiment (Study 3), but nonetheless did greet and introduce participants to the robot. Online interactions allowed our participants to be anonymous to experimenters compared our offline study. Participants can be more sensitive to social norms during in-person experiments [3], e.g., not destroying experimenters' property.

How people hold each other accountable in a shared moral scenario involving a robot can be better understood. Prior research showed that people blamed each other in human-machine teams and not the robot when a robot offers transparent explanations on its mistakes [29]. This showcases how shared responsibility among relevant parties [41] is expected, including the robot, in case it does not transparently explain itself [29]. Yet, the responsibility gap exists between harm done and finding the "right" party to blame [36, 53].

There is more clarity on moral responsibility when harm results due to a robot's agentic action vs. inaction. When a utilitarian robot did divert a trolley, the responsibility for resulting harm was consistently distributed across involved parties— the robot, its designer, and its owner were blamed [31]. But, a robot's deontological *inaction* (a decision to not divert) showed inconsistent blame towards three parties involved [31]. Responsibility for *inaction* is harder to account for. If robots are expected to only make utilitarian decisions [35] and if they are also seen to be responsible, alongside others, when they do make utilitarian decisions [31], distributed responsibility can be a possibility. Yet, robots that act in accordance with other ethical positions, e.g., a deontological decision, are more difficult to factor in. So far the expected norm is that robots will be blamed for allowing passive harm, but not for causing active harm to save others. The norms are indeed different for humans and robots if robots are expected to make utilitarian decisions unlike humans who are expected to *mostly* make deontological decisions.

Given the results of Studies 1 and 2, what is novel to consider is if, why, and how robots should be punished for passively *allowing* harm. Retributive punishment requires some level of institutional coordination and standards [4, 14, 17], which we do not have for robots or non-human agents (as of now). Our participants online did show a tendency to administer punishment, even though the point of punishment is unclear if robots cannot suffer the consequences of physical or psychological punishment like humans can. Robots do not have the potential to know or feel the consequences of their actions like humans do. Yet, this "competence without comprehension" may evolve towards comprehension [11, 12] with more complex AI.

Perhaps retributive punishment towards an artificial agent "expresses *our* disappointment or anger" [50, p. 103] at the wrongdoer in a structural, systemic way because a robot cannot suffer physically or psychologically like us. A robot's lack of emotional displays when it committed an act that is considered blameworthy or punishment-worthy could trigger our reactive attitudes like justified anger, even if a moral patient one feels anger on behalf of is a fictional, anonymous miner. People may reasonably know that blaming or punishing a robot may not do much. But when there is nowhere or no one to direct our reactive attitudes like blame towards (due to responsibility and retributive gaps) people may not be able to practice communicative blame for fostering understanding, repentance and promises of reform [18]. Then, people may seek out institutionalized practices of punishment. One danger would

be moral scapegoating [7] to find something to hold responsible, even a robot.

Perhaps a robot's transparent explanation on what happened [29], coupled with appropriate artificial emotions as our emotional robot displayed, could ameliorate the need for punishment that people want to administer. One connection is that retributive justice regarding psychopaths also considers legal liability as a form of punishment, even if psychopaths may be immune to blame or feeling the gravity of directed blame [20]. An argument could be made that psychopaths also have moral "competence without comprehension", to borrow Dennett's phrase (2009, 2017). Holding psychopaths morally accountable, even if they may be morally "color-blind" is more about our standards of societal justice and ways to direct our feelings of injustice (since directing them to psychopaths or robots is not optimal).

When the weight of interpersonal blame cannot serve its function for regulating interpersonal moral norms and boundaries [48], we may turn to social institutions. If and how institutional practices like retributive punishment can apply to highly autonomous robots is far from clear. If the retributive gap is concerning, robots may as well be treated similarly to psychopath to account for our moral outrage or justified anger in an institutional framework. Or, artificial agents' emotional displays should be more seriously adopted, so that they can at least apologize, admit to mistakes, or act emotionally burdened when there can be no singularly right moral decision to take during "best of possible evils" scenarios when harm towards a person or people by an autonomous system occurs.

Many future paths can be taken. A broader set of participants can help, such as considering people from different culture or ethnic groups, gender-based sampling, and socio-economic status, which were not the focus of our current research. The how and why behind people's likelihood to punish, but not blame a robot, based on its emotional displays (or its lack thereof) requires further research. More studies that look at both online and offline environments to study the same constructs, e.g., blame or punishment, are needed. Perhaps the main distinction is that relevant prior literature on this topic consist of surveys that portray the moral scenario in third-person [31, 35], not based on first-person interaction with a robot in online or offline environments. Thus, there are many intersections that future research can explore: online vs. offline environments, survey vs. direct interaction, third-person observer vs. first-person interactant, presence of humans vs. none, anonymous interaction vs. non-anonymous interaction, and blame vs. punishment. Lastly, scenarios chosen are important. While we deployed the trolley dilemma, a greater variety of morally loaded situations would add depth to future research.

7 CONCLUSION

Our three studies were on whether or not a robot's artificial emotions and perceived mind affects people's likelihood to blame or punish it for passively allowing a person to die (hypothetically) to save more lives. We found no support for the effect of perceived emotions on people's desire to punish or blame a robot *in person*. But in two online studies, people were willing to punish, but not blame, a robot. The robot's lack of perceived patiency (capacity to feel) is a possible reason why people may punish a robot (Figure 3), though people consistently perceived greater agency than patiency in a robot, even if it behaved emotionally (Figure 5).

There are interesting societal implications that stem from our studies on people's moral expectations toward robots. In particular, an open consideration is on if and how robots should be incorporated as a part of our justice system. When real tragedies involving robots strike and no person is (or feels that they are) truly at fault for causing human deaths, our need to assign blame or punishment may go unmet due to responsibility and retributive gaps. But, whether it is morally advisable to have artificial scapegoats and carriers of bad news is uncertain. Further, if people are not willing to blame robots, but potentially willing punish them, what the future justice system would look like to accommodate this is unclear. There other issues that are worthy of deeper investigations. Open debates are on whether robots should indeed be punished, what punishing robots consists of, if our anonymity matters in punishing robots, for whom robots should be punished (if at all), and what larger impact punishing artificial agents can have on humans should be examined.

In the human world, repenting for potential sins or perceived wrongs have never been easy in ethical gray zones. Robots will fare no better, whether we attribute some moral status or mind-related traits to them or not. Though robots are far from perfect, their artificial commiseration and emotions that *seem* real is an option to address our real, hurt feelings when no particular people can be responsible. Due to the potential responsibility and retributive gaps, artificial moral emotions of an artificial scapegoat may be more ameliorating than the absence of real emotions, understanding, responsibility, and remorse in humans who may remain legally and morally unaccountable for victims' outrage, anger, and sense of injustice. These reactions and feelings may deserve to be recognized, be it by human or artificial beings.

8 ACKNOWLEDGMENTS

The authors thank the reviewers, as well as Eline Cloudt, Maxine Derksen, Anne Kok, and Scott Elliot for their involvement and insightful discussions.

REFERENCES

[1] Arthur Aron, Elaine N Aron, and Danny Smollan. 1992. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology* 63, 4 (1992), 596–612.
[2] Christoph Bartneck and Merel Keijsers. 2020. The morality of abusing a robot. *Paladyn, Journal of Behavioral Robotics* 11, 1 (2020), 271–283.
[3] Iris Bohnet and Bruno S Frey. 1999. Social distance and other-regarding behavior in dictator games: Comment. *American Economic Review* 89, 1 (1999), 335–339.
[4] Kevin M Carlsmith, John M Darley, and Paul H Robinson. 2002. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* 83, 2 (2002), 284.
[5] Andreas Brekke Carlsson. 2017. Blameworthiness as deserved guilt. *The Journal of Ethics* 21, 1 (2017), 89–115.
[6] Mark Coeckelbergh. 2009. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society* 24, 2 (2009), 181–189.
[7] John Danaher. 2016. Robots, law and the retribution gap. *Ethics and Information Technology* 18, 4 (2016), 299–309.
[8] John Danaher. 2019. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics* (2019), 1–27.
[9] Stephen Darwall. 2004. Respect and the second-person standpoint. In *Proceedings and Addresses of the American Philosophical Association*, Vol. 78. JSTOR, 43–59.
[10] Celso M de Melo and Jonathan Gratch. 2015. People show envy, not guilt, when making decisions with machines. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 315–321.

[11] Daniel Dennett. 2009. Darwin's "strange inversion of reasoning". *Proceedings of the National Academy of Sciences* 106, Supplement 1 (2009), 10061–10065.

[12] Daniel C Dennett. 2017. *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company.

[13] Antony Duff and Zachary Hoskins. 2017. Legal Punishment. *Stanford Encyclopedia of Philosophy* (Jul 2017). https://plato.stanford.edu/entries/legal-punishment/#PosRetMeaDes (Accessed on 12/23/2020).

[14] Robin Antony Duff. 2003. Probation, punishment and restorative justice: Should Al Turism be engaged in punishment? *The Howard Journal of Criminal Justice* 42, 2 (2003), 181–197.

[15] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and Machines* 14, 3 (2004), 349–379.

[16] Philippa Foot. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5 (1967).

[17] Michel Foucault. 2012 [1975]. *Discipline and punish: The birth of the prison*. Vintage.

[18] Miranda Fricker. 2016. What's the point of blame? A paradigm based explanation. *Noûs* 50, 1 (2016), 165–183.

[19] Caleb Furlough, Thomas Stokes, and Douglas J Gillan. 2019. Attributing Blame to Robots: I. The Influence of Robot Autonomy. *Human Factors* (2019), 1–11.

[20] Geoffrey P Goodwin, Jared Piazza, and Paul Rozin. 2014. Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology* 106, 1 (2014), 148–168.

[21] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of mind perception. *science* 315, 5812 (2007), 619–619.

[22] Kurt Gray, Chelsea Schein, and Adrian F Ward. 2014. The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General* 143, 4 (2014), 1600.

[23] Kurt Gray, Liane Young, and Adam Waytz. 2012. Mind perception is the essence of morality. *Psychological inquiry* 23, 2 (2012), 101–124.

[24] Jonathan Haidt. 2003. The moral emotions. In *Series in Affective Science. Handbook of Affective Sciences*, H. H. Goldsmit R. J. Davidson, K. R. Scherer (Ed.). Vol. 11. Oxford University Press, 852–870.

[25] Francis Hutcheson. 2008 [1725]. *An Inquiry into the Original of Our Ideas of Beauty and Virtue: In Two Treatises*. Liberty Fund.

[26] David O Johnson and Raymond H Cuijpers. 2019. Investigating the effect of a humanoid robot's head position on imitating human emotions. *International Journal of Social Robotics* 11, 1 (2019), 65–74.

[27] Merel Keijsers and Christoph Bartneck. 2018. Mindless Robots Get Bullied. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 205–214.

[28] Mansur Khamitov, Jeff D Rotman, and Jared Piazza. 2016. Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition* 146 (2016), 33–47.

[29] Taemie Kim and Pamela Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 80–85.

[30] Joshua Knobe. 2003. Intentional action and side effects in ordinary language. *Analysis* 63, 3 (2003), 190–194.

[31] Takanori Komatsu. 2016. How do people judge moral wrongness in a robot and in its designers and owners regarding the consequences of the robot's behaviors?. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1168–1171.

[32] Minha Lee, Gale Lucas, and Jonathan Gratch. 2021. Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator and ultimatum games. *Journal of Multimodal User Interfaces*. In press (2021), 1–15.

[33] Minha Lee, Gale Lucas, Johnathan Mell, Emmanuel Johnson, and Jonathan Gratch. 2019. What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 38–45.

[34] Daniel T Levin, Stephen S Killingsworth, Megan M Saylor, Stephen M Gordon, and Kazuhiko Kawamura. 2013. Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people. *Human–Computer Interaction* 28, 2 (2013), 161–191.

[35] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 117–124.

[36] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6, 3 (2004), 175–183.

[37] Daniel McDermott. 2001. The permissibility of punishment. *Law and Philosophy* 20, 4 (2001), 403–432.

[38] Robert M McFatter. 1978. Sentencing strategies and justice: Effects of punishment philosophy on sentencing decisions. *Journal of Personality and Social Psychology* 36, 12 (1978), 1490–1500.

[39] Catrin Misselhorn. 2015. Collective agency and cooperation in natural and artificial systems. In *Collective Agency and Cooperation in Natural and Artificial Systems*, Catrin Misselhorn (Ed.). Springer, 3–24.

[40] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Conference Companion on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 204.

[41] Sven Nyholm. 2018. Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics* 24, 4 (2018), 1201–1219.

[42] Sven Nyholm. 2020. *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.

[43] Yohsuke Ohtsubo. 2007. Perceived intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect. *Japanese Psychological Research* 49, 2 (2007), 100–110.

[44] Jesse Prinz. 2008. Is morality innate? In *Moral psychology. The evolution of morality: Adaptations and innateness*, Walter Sinnott-Armstrong (Ed.). Vol. 1. MIT Press Cambridge, 367–406.

[45] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

[46] Beat Rossmy, Sarah Theres Völkel, Elias Naphausen, Patricia Kimm, Alexander Wiethoff, and Andreas Muxel. 2020. Punishable AI: Examining Users' Attitude Towards Robot Punishment. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, New York, NY, USA, 179–191.

[47] P. Salvini, G. Ciaravella, W. Yu, G. Ferri, A. Manzi, B. Mazzolai, C. Laschi, S. R. Oh, and P. Dario. 2010. How safe are service robots in urban environments? Bullying a robot. In *19th International Symposium in Robot and Human Interactive Communication (ROMAN)*. 1–7.

[48] Thomas Michael Scanlon. 2013. Interpreting blame. In *Blame. Its nature and norms*, D. Justin Coates and Neal A. Tognazzini (Eds.). 84–99.

[49] Simone Schnall, Jonathan Haidt, Gerald L Clore, and Alexander H Jordan. 2008. Disgust as embodied moral judgment. *Personality and social psychology bulletin* 34, 8 (2008), 1096–1109.

[50] David Shoemaker. 2013. Blame and punishment. *Blame: Its nature and norms* (2013), 100–118.

[51] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 219–226.

[52] Eva EA Skoe, Nancy Eisenberg, and Amanda Cumberland. 2002. The role of reported emotion in real-life and hypothetical moral dilemmas. *Personality and Social Psychology Bulletin* 28, 7 (2002), 962–973.

[53] Robert Sparrow. 2007. Killer robots. *Journal of Applied Philosophy* 24, 1 (2007), 62–77.

[54] Peter Frederick Strawson. 2008 [1963]. *Freedom and resentment and other essays*. Routledge.

[55] Judith Jarvis Thomson. 1976. Killing, letting die, and the trolley problem. *The Monist* 59, 2 (1976), 204–217.

[56] John Voiklis, Boyoung Kim, Corey Cusimano, and Bertram F Malle. 2016. Moral judgments of human vs. robot agents. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 775–780.