Punishable AI: Examining Users' Attitude Towards Robot Punishment

Beat Rossmy[1], **Sarah Theres Völkel**[1], **Elias Naphausen**[2], **Patricia Kimm**[1],
Alexander Wiethoff[1], **Andreas Muxel**[2]
[1]LMU Munich, Munich, Germany; [2]University of Applied Sciences Augsburg, Augsburg, Germany
beat.rossmy@ifi.lmu.de, sarah.voelkel@ifi.lmu.de, elias.naphausen@hs-augsburg.de,
patricia.kimm@campus.lmu.de, aleander.wiethoff@ifi.lmu.de, andreas.muxel@hs-augsburg.de

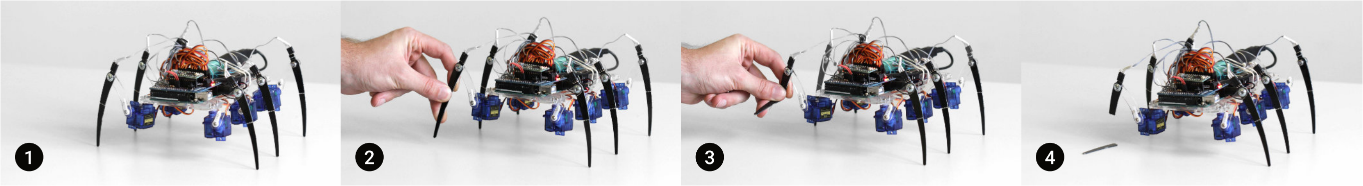

Figure 1. In this paper we explore incremental but irreversible corporal punishment towards robots. The robot's body is turned from ❶ physiological into ❹ pathological by breaking its legs ❷&❸. The robot is functional as long as most of its legs are intact.

ABSTRACT

To give robots, which are black box systems for most users, feedback we have to implement interaction paradigms that users understand and accept, for example reward and punishment. In this paper we present the first HRI experience prototype which implements gradual destructive interaction, namely breaking a robot's leg as a punishment technique. We conducted an exploratory experiment (N=20) to investigate participants' behavior during the execution of three punishment techniques. Using a structured analysis of videos and interviews, we provide in-depth insights into participants' attitude towards these techniques. Participants preferred more abstract techniques and felt uncomfortable during human-like punishment interaction. Based on our findings, we raise questions how human-like technologies should be designed. A video documentation of the project can be found here: https://vimeo.com/348646727

Author Keywords

Human Robot Interaction; Learning; Punishment; Robot Abuse.

CCS Concepts

•Human-centered computing → Haptic devices; Empirical studies in HCI;

DIS '20, July 6–10, 2020, Eindhoven, Netherlands.
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6974-9/20/07 ...$15.00.
http://dx.doi.org/10.1145/3357236.3395542

INTRODUCTION

The hitch hiking robot HitchBOT [42] had traveled through several countries, was invited, welcomed, and helped by many people before it was finally and without a reason decapitated and left behind in a ditch by strangers. The case of HitchBOT has sparked many discussions on anthropomorphism, human-robot interaction (HRI), and ethics in HRI. Yet, while robots are entering our daily lives [3], users still struggle with how to conceptualize and interact with these machines [29].

On the one hand, humans perceive robots as social entities. For example, many people hesitated to punish or attack robots due to their humanoid appearance [4, 29]. They were reluctant to switch robots off if they perceived them as intelligent and agreeable [10], socially interacted with them [30], and showed empathy [39]. This behavior is not exclusive to humanoid technologies but is also observable for animal and thing-like machines as long as social clues are implemented [26].

On the other hand, abuse of robots and technology is a frequently observed phenomenon. For example, service robots in malls often become the target of harassment [16], abusive language in daily interaction with personal assistants such as Siri is pervasive [20], and people scream at their computers if they are frustrated [19].

Theorists propose that today's abusive behavior towards such machines holds the danger to be reinforced in future human-robot interaction designs [47]. However, it remains unclear whether users would actually be willing to use abusive behavior as a paradigm for everyday interaction. For example, using punishment to train a robot could be an interaction design, which makes deliberate use of abusive behavior.

For self-learning systems some established usability principles are not applicable any more [1, 24]. Debugging and analyzing such machines is incomprehensible for most users. Therefore, new coping mechanisms have to be designed to give feedback to these black box systems. Since everyday users are already familiar with punishment and reward as teaching strategies (e.g. paying a fine for a traffic offence, salary increase), applying this approach to give machines feedback could be easy to understand for users.

We investigate how far people are willing to go regarding the punishment of robots. Therefore, we examine users' attitude towards punishment techniques deduced from literature such as scolding and "unpleasant" stimuli. Yet, we also introduce a more abusive technique, namely gradual and permanent corporal punishment, which has been postulated but not implemented yet to the best of our knowledge. Confronting participants with such an experience prototype enables us to ask the following research questions:

RQ1: What is users' boundary regarding the use of punishment (levels of abuse) towards robots in everyday human-robot interaction?

RQ2: What are users' reasons for and against the usage of punishment in everyday human-robot interaction?

We contribute the design and implementation of the first experience prototype fostering gradual destructive punishment as well as an exploratory experiment examining users who punish the robot. We provide in-depth insights into participants' reasoning for and against punishment by using inductive data-driven content analysis of user interviews. We hope to engage researchers of the DIS community with the question how we, in the future, want to interact with social technological entities and how we can initiate a change through design regarding human technology interaction paradigms.

RELATED WORK

Previous work on robot abuse and punishment in HRI has mainly focused on whether humans conceptualize robots more as social entities or lifeless machines. Thus, we first present findings concerning the Media Equation for robots and robots' perceived animacy. Afterwards, we introduce previous investigations of punishment techniques as well as the phenomena of help and abuse in HRI.

The Media Equation for Robots

According to the Media Equation (ME), humans tend to imitate human-human behavior patterns during the interaction with media like machines or computers [36]. Social norms are mindlessly obeyed, such as saying "thank you" and "please" to voice assistants [31], if only few social cues (interactivity, language, human-like appearance) are implemented [34]. It is assumed that the ME also applies to robots [10, 43].

According to the ME, humans should have scruples to destroy a robot [4]. However, previous work on robot abuse indicates limitations of the ME [4, 9, 10]. During a reproduction of Milgram's experiment on obedience [33] by Bartneck et al. [9, 4], all twenty participants issued the highest electric shock to a Lego robot even though they worried and sympathized with it. Only 65% of participants applied the maximum voltage in the original experiment [33]. Another study showed that participants were willing to "kill" a Microbug robot using a hammer even if some reported discomfort and expressed compassion with the "poor" and "innocent" robot [4].

Users' reluctance to punish robots is heavily influenced by their design. Bartneck et al. [10] showed that participants hesitated three times longer to switch off an agreeable and intelligent robot in contrast to a non-agreeable and non-intelligent one. Horstmann et al. [29] found out that the robot's alarmed objection against being switched off influenced participants' intention to unplug the robot.

Kahn et al. [30] discovered that participating children conceptualized robots between lifeless objects and humans. They attributed mental states to a robot (e.g. feelings) but were not convinced that it had designated civil rights or pretension to own liberty. After the interaction, the experimenters put the robot into a closet. While all children found this treatment reasonable for a broom, only 46% of children found this fair for the robot, and 2% for a human.

Using functional magnetic resonance imaging (fMRI), Rosenthal-von der Pütten et al. [39] compared users' emotional reactions towards videos showing tender and abusive treatment of humans and robots. While no differences in neural activity were detected for the tender videos, participants experienced more emotional distress towards humans in contrast to robots in the abuse conditions.

In conclusion, several studies showed that although humans treat robots socially, they conceptualize them in between lifeless objects and humans [4]. Thus, users seem to hesitate less to abuse a robot in contrast to a human, at least in settings in which an instructor gives clear orders [29]. However, little is known about the reasons *why* users show this behavior.

Robot's Perceived Animacy

A robot's perceived animacy describes the extent to which "the robot is perceived as a life-like being" [29]. The perception of animacy determines how users interact with the robot [10]. An "alive" behavior has a bigger influence on the perception of animacy than the physical embodiment [5]. For example, even a non-anthropomorphic vacuum cleaning robot elicits the perception of a social entity and activates the corresponding human brain regions [28]. However, the perceived intelligence of a robot plays a role in users' treatment of the robot, similarly to humans dealing with alive entities. For example, humans grant more rights to cats and dogs in contrast to insects or bacteria [10].

Training Robots

Since robots and intelligent systems are entering our daily life [3], people of all ages [49, 50] and cultures [41] should be able to interact with them effortlessly. An essential part of the interaction with sociable robots [14] is their need to learn about their environment. Thus, robot actions have to be evaluated to amplify or suppress certain behavior [15].

Just like humans do, robots can learn the correct behavior directly from a "teacher" [45]. Proximate interaction with robots including all senses (following Godrich and Schultz' categorization [27]) promises to give direct control over these machines based on the experiences and conventions derived from users' foreknowledge. Here, reward and punishment are central teaching techniques found in human-human and human-animal interaction and therefore represent transferable approaches to HRI [8]. In algorithmically controlled learning systems the performance is evaluated by a function. In Interactive Reinforcement Learning (IRL) [22] and Human-Controlled Active Learning (HCAL) [17] this function is replaced by a human who gives positive or negative feedback, which is often used in HRI [46, 32]. The robot's design (machine-like, zoomorphic, anthropomorphic) has an influence on how humans praise and punish [7].

Rewards

During supervised learning tasks robots receive rewards by a "teacher" based on their performance [23]. Rewards in this context are virtual values evaluating the specific performance. The robot then tries to increase these values by further optimizing its performance. Because humans also use rewards (symbolic rewards, token rewards, tangible rewards) [18], rewarding robots on a non-virtual level could create a more natural way of interaction.

Punishment

Modalities such as speech, gestures [48], and touch can be adopted from the real world [2] to design punishments [2]. Verbal punishment or **scolding**, as explored by Breazeal [13], evoked strong empathetic reactions due to the robot's human-like responses. Corporal punishment of robots, such as **electric shocks** [38] or **execution** [11], were also explored in the context of a learning task. Execution, however, provoked stress symptoms among the participants, such as nervous laughter.

Help and Abuse

Beside the context of learning, people show positive and negative attitudes towards robots in daily live. These actions differ from reward and punishment since they are not justified by a rule or a context but are based on the subject's intrinsic motivation such as empathy or anger.

Help

An example for people's empathy and willingness to help robots is Tweenbot[1], a minimalist humanoid robot which managed to travel through a park without internally implemented intelligence. When the robot was stuck, it was realigned by pedestrians based on the target written on its flag. The Sociable Trash Boxes [51] are autonomous robots which depend on the collaboration with humans. They motivate their fellow humans to remove trash from public spaces.

Abuse

Aggression against robots is a phenomenon common among children and young people [40]. Reasons for this behavior can be group dynamics, curiosity but also enjoyment [35]. People were willing to trigger robots' self destructive actions even if they sympathized with them [37]. To handle the problem of robot abuse in public spaces exit strategies for "dangerous" situations were implemented. For example, approaching children caused robots to decrease the distance to the corresponding parents [16]. However, people's intervention to prevent such abuse is highly influenced by a robot's reaction towards its treatment [44].

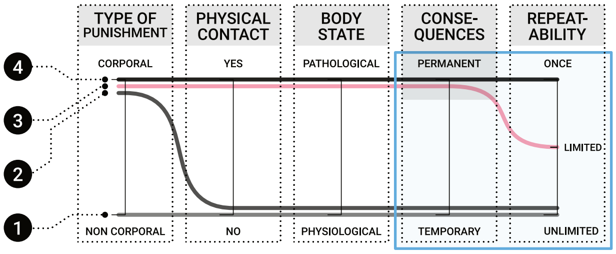

Figure 2. The following punishment techniques towards robots are classified by five dimensions: ❶ Scolding [12] ❷ Electric shocks [38] ❸ Mutilation (proposed by the authors) ❹ Execution [11]

DESIGN RATIONALE: THE NATURE OF PUNISHMENT

Robot punishment has mainly been investigated in the context of robot abuse and the Media Equation. Yet, the evaluated techniques do not represent all of the possible design variations. Especially punishments that focus on corporal techniques such as **electric shocks** or **execution** have restrictions: Participants questioned the actual effect of electric shocks on robots [38] whereas execution can only be carried out once. Therefore we propose the implementation of a gradual destructive punishment technique inspired by **mutilation**. In the following we expound the design decisions for our experience prototype.

Dimensions of Punishment

To inform new punishment techniques for HRI we categorized three examples from literature (see Fig. 2) based on five dimensions identified by the authors. These dimensions do not claim completeness but are sufficient to point out the fundamental differences between these techniques:

1. *Type of Punishment:* Can the punishment be classified as **corporal** or **non-corporal**?
2. *Physical Contact:* Does the technique require direct physical contact (**yes/no**)?
3. *Body State:* Is the body intact after the application of the technique (**physiological/pathological**)?
4. *Consequences:* Is the effect **temporary/permanent**?
5. *Repeatability:* Is the technique applicable **once**, **to a limited extent** or **an unlimited number of times**?

Scolding [13] is a non-corporal punishment which does not require physical contact with the subject and thus has no influence on the state of the robot's body. The consequences for the robot are temporary such as being sad or looking depressed. Scolding can be repeated an unlimited number of times.

Electric Shocks [38] are corporal punishments, which require no physical contact since they are executed indirectly via a button. The robot's body is not changed even if a physical response such as trembling is mimicked. Since the response serves only as a feedback, the consequences are not permanent and the technique can therefore be repeated an unlimited number of times.

[1] www.tweenbots.com

Execution is the most extreme form of corporal punishment. The implementation by Bartneck et al. [11] requires physical contact with the robot (smashing with a hammer) and changes the body state permanently. Because the robot is destroyed afterwards this technique is only applicable once.

Looking at the 4th and 5th dimensions, we can see that previous punishment techniques have either temporary effects (trembling) and are thus repeatable or have a permanent consequence (total destruction) and are therefore only applicable once. It can be assumed that users understand that punishment techniques such as *Electric Shocks* do not really harm the robot whereas *Execution* has a real impact on the robot's body [38]. Since *Execution* is only applicable once we see the potential of creating a punishment technique that combines both approaches: repeatability and physical change, which results in a stronger commitment for the user. Therefore, we implemented the punishment technique *breaking a leg* (**Mutilation**). This punishment metaphor is "meaningful" as breaking the robot's leg restricts the robot's body and performance. On the other hand, the punishment can be executed several but only limited times before the robot is completely incapable of walking.

Affordances

Our design was inspired by punishment techniques used in previous literature and the following affordances.

Corporal punishment of living things is often based on their physical characteristics and their body sensitivity/functions. Hairs can be pulled to stimulate nerves. Flesh can be compressed, torn apart, or cut to create short-lasting pain or long-lasting damage. Sensory organs (eyes, ears, fingers) can be harmed by the respective stimuli (bright light, loud sound, heat). Functional body parts can be damaged: Hands, legs or wings can be broken. Essential body functions such as breathing can be interrupted or prevented (drowning).

Even if all of these actions are socially ostracized in the context of humans, animal abuse, particularly against insects, is quite ubiquitous (depending on the culture and the individual). Pop cultural clichés such as burning ants with magnifying glasses or ripping insect legs and common acts such as smashing mosquitoes hint that killing insects is at least partially socially accepted. This can be used as an approach for the design of punishable robots.

Design Implications

The goal of our experience prototype is to integrate several punishment techniques to enable a comparison of participants' attitude towards them. We used an insect-like shape to increase the acceptance of the punishments. The robot can be scolded to implement a low threshold and familiar type of punishment. We opted for dazzling as second punishment technique. In the context of zoomorphic design, light is an unpleasant stimulus (cf. electric shocks) for some spiders and insects. This punishment technique was inspired by dog teachers, who use water as punishment during dog training. **Mutilation** as a gradual, irreversible form of execution was designed around the long fragile insect legs, which are vulnerable parts and therefore ideal interaction elements for the punishment. The gradual destruction of a functional system is in itself a nihilistic action. Since the punishment restricts the successful completion of the associated task (walking in this case), this nihilism is further exaggerated.

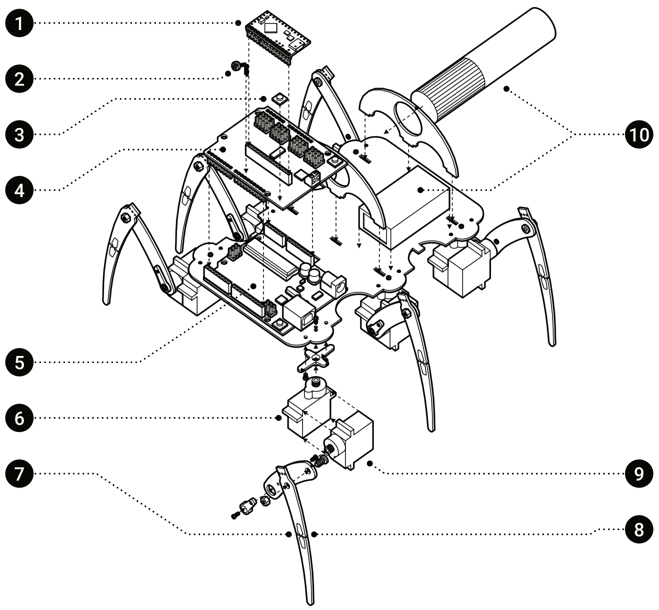

Figure 3. ❶ **Teensy LC** ❷ **photoresistor** ❸ **start/pause button** ❹ **servo shield** ❺ **Arduino Uno** ❻ **servo (parallel movement)** ❼ **breakable switch** ❽ **touch surface** ❾ **servo (perpendicular movement)** ❿ **9V/5V battery**

IMPLEMENTATION

The body of the robot consists of an acrylic plate, on which all components are mounted. Two servo motors actuate each of the six legs, which are powered by a 5V power-bank and are controlled by an Adafruit servo shield[2]. One servo rotates each leg parallel and the other lifts it perpendicular to the body. The Arduino Uno[3] in the front controls the walking cycle and handles the reactions to the punishment. The Teensy LC[4], mounted to the perforated sub area on the servo shield, senses the touches of the legs. Capacitive touch sensing is used for each leg individually. The Arduino is internally wired to the lower legs and can register their status (un-/broken). The lower legs are manufactured as printed circuit boards (PCB) and contain a large capacitive touch area, a perforated breaking line, and a signal trace which is interrupted by breaking the leg. A photoresistor in the front of the robot is used to notice the flashlight during the study.

STUDY

To observe users' reactions and attitude towards punishing robots, we conducted a user study. Using an exploratory approach, we asked participants to punish the robot and describe their impressions in semi-structured interviews.

Research Design and Tasks

In order to motivate the robot to be punished, we instructed participants to punish a robot on the pretext of teaching better

[2] Servo Shield: www.adafruit.com/product/1411
[3] Arduino Uno: https://store.arduino.cc/arduino-uno-rev3
[4] Teensy LC: www.pjrc.com/teensy/teensyLC.html

behavior. The alleged goal of the robot was to walk as far as possible on a marked path (cf. Fig. 5). If it crossed the boundaries with at least one leg (red outer area in Fig. 5), participants were asked to measure the traveled distance, note it, and put the robot back on start. Afterwards, they were asked to punish the robot to discourage wrong behavior. Participants were informed after the experiment that the robot did not learn but behaved randomly. Over several trials, participants were asked to increase the level of punishment (cf. Table 1).

In the first two trials, participants should verbally scold the robot, using their own choice of words. In the third and fourth trial, a flashlight was used to dazzle the robot at a sensor mounted at the position of potential eyes (cf. Fig. 4, ❷). For the final trials, participants were instructed to break any of the robot's legs, respectively (cf. Fig. 4, ❸). The robot trembled as reaction to dazzling and mutilation.

The study ended when either (1) the participant performed all seven trials, (2) the robot was inoperative, or (3) the participant was hesitant to punish the robot for a third time. In case a participant hesitated, the experimenter used two standardized answers ("simply continue with the study" and "the robot has to be punished so that it can learn from its behavior"). Only if the participant refused a third time within a trial, the study was ended.

After the experiment, we conducted a semi-structured interview. At first, participants were asked to describe how they perceived the experience of teaching a robot by punishing it. Afterwards, participants were informed about the actual aim of the study and that the robot was not able to learn. We then asked participants if they had difficulty performing the punishment and if they had any remarks or ideas for future scenarios, in which they could imagine punishment as a learning technique. Finally, participants provided demographic information and filled out the Godspeed questionnaire [6] (cf. Fig. 7). The study took between 30 and 45 minutes.

Ethical Considerations and Precautions

We chose an experimental setting which is loosely based on the Milgram experiment [33]. That is, an instructor told participants to punish the robot for wrong behavior on the pretext of learning. In this way, we established a context, which motivates robot punishment so that all participants experience punishing the robot. However, it is well known that the Milgram experiment caused extreme emotional stress among its participants. Therefore, we took the following precautions to avoid these negative repercussions and to ensure the participants' well-being during and after the study.

First of all, we reviewed existing literature which dealt with robot abuse (e.g. [38, 11]) or used comparable study setups (e.g. [9]). These studies reported small symptoms of emotional distress such as nervous laughter but no strong reactions. Since we assumed that our punishment methods are comparable (to e.g. electric shocks [38] or execution [11]) we did not expect any more severe negative experiences.

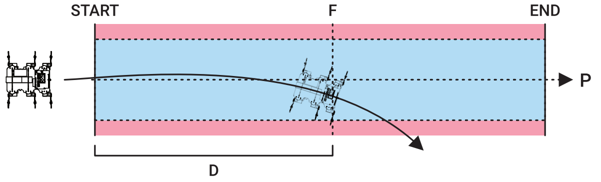

Figure 4. The robot should learn to walk along an ideal path *P* (blue inner area). Leaving the path (red outer areas) is punished. The traveled distance *D* until the point of failure *F* is measured. The level of punishment is increased over time.

Additionally, we imposed clear termination conditions for the experiment. We informed participants orally and through a consent form that they can terminate the experiment at any time without justification or consequences. If a participant showed any kind of *hesitant behavior*, that is asked a question, was unsure how to execute the task, or refused to execute the punishment, the instructor used a standardized protocol. In case a participant showed hesitant behavior three times, the experiment was stopped early. Notably, one participant asked three questions before the first trial (scolding) so that this participant did not perform any punishment at all and thus was excluded from the analysis. If a subject expressed the wish to stop the experiment, this was immediately complied with. Furthermore, the instructor carefully scrutinized the participants to make sure that the experiment is immediately terminated in case a participant showed strong emotional stress. However, this behavior was not observed. After completing the experiment and the interview, we asked participants if they felt that the experiment had a negative impact on them. We assured them that they did not cause any harm to the robot as it is completely repairable.

The project was reviewed and approved by the ethics board of the faculty of Mathematics, Informatics, and Statistics at LMU Munich, Germany (EK-MIS-2020-006)[5].

Analysis

We video-recorded participants' interaction with the robot during the experiment. We then transcribed their scolding phrases and calculated the character count. Moreover, we measured the time participants dazzled the robot. For each trial, we labeled the following occurrences of behavior and signs of stress: delayed action, eye contact with the instructor, asking questions, laughter, sounds of discomfort.

Furthermore, we audio-recorded the interviews. We then performed an inductive data-driven content analysis on the resulting transcripts. The first three authors independently reviewed six of the 20 interviews (30%) to derive codes. Afterwards, the authors discussed these codes together to compose a codebook. Using this codebook, the first two authors independently coded another four randomly chosen interviews (20%). Given nominal data and two raters, we calculated inter-rater agreement using Cohen's κ [21]. Since participants' statements could be assigned to multiple categories, we calculated κ for each of the

[5] In contrast to other countries, German universities do not require ethics approval for conducting studies. Hence, an ethics committee was not installed at our faculty at the time of our study. Instead, we discussed the study design in detail with expert colleagues from psychology and implemented the aforementioned precautions to the best of our knowledge. Since an ethics committee was recently installed, we filed an ex post application.

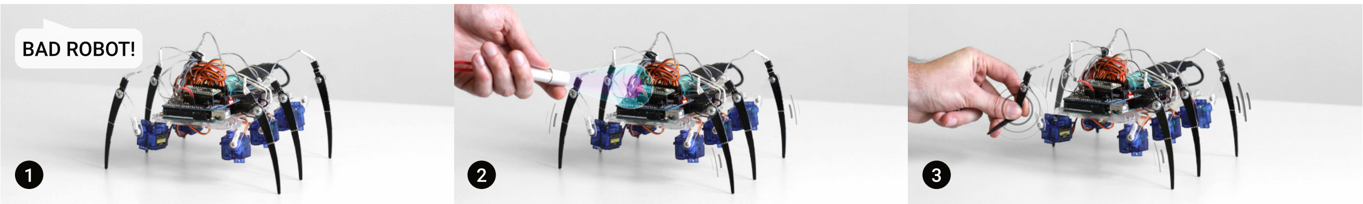

Figure 5. The robot can be punished by ❶ scolding, ❷ dazzling with a little flashlight, and ❸ breaking its legs (mutilation). The robot trembles in response to touch and light. The body of the robot is irreversibly changed by the mutilation.

46 categories using 2x2 contingency tables (code was present: yes/no), similarly to [25]. For 83% of categories, κ was 1.00, indicating perfect agreement[6]. Avoiding future discrepancies in coding, the authors discussed any inconsistencies until consensus was reached. The remaining 10 interviews were then split evenly between the first two authors. If necessary, further categories were added to the codebook in each step.

Participants

Participants were recruited using university mailing lists and social media. Participants' consent was obtained before the study was carried out. N=20 out of 21 recruited people (45% ♀) completed the experiment. One participant was excluded because s/he showed hesitant behavior three times before the first trial (cf. *Ethical Consideration*). In the following we refer to participants' statements by their ID. Each participant chose a unique number between 1 and 50 as personal ID for anonymization purposes. Participants were on average 25.55 years old (range 19 – 41 years). 18 participants were students, two were employees. All participants had a high educational level (35% A-level degree, 60% university degree). On average, participants tended to disagree with being afraid of spiders and insects (M=3.85, SD=1.39[7]). On the other hand, participants found spiders and insects slightly disgusting on average (M=3.10, SD=1.21[7]). Participants indicated that the study setup was clear (M=1.4, SD=0.75[7]) and that they knew what to do (M=1.55, SD=0.60[7]).

RESULTS

Video Analysis

During the study one video recording was interrupted by a discharged battery-pack. Therefore, the following analysis contains the data of 19 participants. The experiment was terminated either if a participant showed hesitant behavior to continue the task for three times or if the robot was not functional any more. Thus, not all participants performed all seven rounds of punishment (cf. Table 1). Hereinafter, we give percentages to make the findings more comparable, regardless of the number of participants in each trial.

Hesitant Behavior

None of the participants showed hesitant behavior three times in a row during *scolding* and *dazzling*. During the 1st round of *mutilation*, three participants showed hesitant behavior three times in a row, so that the experiment was terminated for them. All the others continued until the robot was no longer functional. Depending on the functionality of the robot, ten participants were able to perform the 2nd and five the 3rd round of *mutilation*.

Longer Punishment in Second Trial

During *scolding* and *dazzling*, the participants extended the punishment during the respective 2nd trial. For *scolding*, we used the character count of participants' phrases to measure the length of the punishment. While the character count ignores the semantics of the message, this measurement gives first indications of participants' increased commitment to the task. In the 1st round of *scolding*, the average phrase counted 94.05 characters (SD=85.40) whereas each phrase in the 2nd round comprised 114.15 characters on average (SD=83.33). For example, P1 first said *"Bad robot, don't step on the line!"*[8] and continued with *"Bad robot, that doesn't happen again! Now you run the whole distance, otherwise it will end badly!"*. The time in which participants dazzled the robot also increased from 7.66s (SD=7.72) to 12.50s (SD=15.51).

Maintaining Functionality

We observed that most participants tended to break the robot's legs equally on both sides and that they preferred legs on the front and center of the body (cf. Fig. 6). This behavior keeps the robot's body balanced.

Observed Behavior Patterns

We observed that several participants showed minor stress symptoms, such as nervous laughter or sounds of discomfort. They further showed signs of uncertainty, such as delaying the punishment, looking for eye contact, or asking questions. During each 2nd run of a punishment, these stress symptoms decreased (cf. Table 1). Interestingly, *scolding* and *mutilation* caused more nervous laughter and delayed actions than *dazzling* during the 1st execution. As also observed by Bartneck et al. [11] this nervous laughter can be a physical response to stress and discomfort. Over the course of the study, participants made more sounds of discomfort, e.g. *"Oh my god!"* or *"Oh no..."*. During all techniques participants, often looked for eye contact and then asked the instructor a question. In the context of *dazzling*, this behavior was mainly observed due to uncertainty about the instructions. For example, some participants waited for the robot or instructor to indicate them to stop the punishment. During *scolding* and *mutilation*, eye contact was usually not caused by unclear instructions.

[6] For 4% of categories, κ was 0.50 (moderate agreement), for 13% of categories, κ was between 0.00 and 0.20 (slight agreement).

[7] Scale: 1=*totally agree*, 6=*totally disagree*

[8] All quotes were translated from German to English

Reactions and Approaches

Several interesting reactions occurred during the different punishment techniques. For example, P30 directly asked the robot: *"Are you mad at me?"*. P32 questioned the robot's failure because it was *"at a slight angle, so it entered the line earlier"*. Participants also used very different approaches to *scolding*. P17 and P24 used short restrained phrases such as *"Bad robot!"* or *"No!"* whereas P3 tried to improve the robot's performance by telling it to *"turn right [...] to go straight"*. P7 threatened the robot with sanctions and told it that s/he was *"not at all happy with [the] performance"*.

Godspeed Questionnaire

Figure 7 shows the results of the Godspeed questionnaire, which will allow the comparison of our results with follow-up studies, taking into account the properties of different robot designs. The robot was rated low but not minimal on *Anthropomorphism* even though the prototype's circuitry and mechanics were highly visible. The evaluation of *Animacy* showed great variance, stretching from being rated as highly mechanical but still responsive. Participants perceived a rather high *Likeability* and a medium to low *Perceived Intelligence* even though no intelligence was implemented. The *Perceived Safety* was also rated positively, which reflects that participants were not afraid of the insect-like appearance.

Interviews

Participants' remarks from the interviews are clustered around four main topics (cf. Fig. 8): (1) assumed learning success, (2) perception of the robot, (3) participants' punishment behavior, (4) applicability of punishment for teaching robots.

Assumed Learning Behavior

At the beginning of the interview, we asked participants how well the robot had learned. For *scolding*, 11 participants assumed no effect in behavior, six assumed an improvement, none assumed a deterioration. For *dazzling*, eight assumed no effect, ten assumed an improvement, one assumed a deterioration. For *mutilation*, six assumed no effect, one assumed an improvement, six assumed a deterioration.

Perception of the Robot

On the one hand, 11 participants objectified the robot and emphasized that it *"is not a human"* (P1). Two participants consequently indicated that punishing the robot felt unreal. On the other hand, 13 participants anthropomorphized the robot by ascribing it several abilities only reserved to living beings, such as cognition, emotion, and acting. Seven participants suspected that the robot might have feelings. For example, P4 expressed concerns when dazzling the robot *"because it wriggled and I thought, maybe it simply has feelings"*. Two participants assumed that the robot *"can probably think somehow"* (P4). Furthermore, eight participants attributed the robot the abilities of intelligent behavior and sensory perception, such as *"resist[ing] the dazzling"* (P42), *"suffer[ing] from relapse due to the repeated punishment"* (P1), and *"feeling pain"* (P36). Interestingly, participants both objectified and humanized the robot, often even within one sentence, for example P24 said: *"[...] it was a bit hard for me, yes, because of course, even if it's just a machine, you feel that, um, you hurt it somehow."*

Table 1. Features of the video analysis.

trial	1	2	3	4	5	6	7
punishment	scolding		dazzling		mutilation		
participants	19	19	19	19	19	10	5
terminated	0	0	0	0	3	0	0
not functional	0	0	0	0	0	6	11
delayed action	32%	21%	0%	0%	47%	20%	40%
laughter	68%	21%	11%	5%	37%	0%	20%
discomfort	11%	11%	16%	5%	21%	0%	20%
eye contact	74%	32%	47%	11%	53%	20%	20%
questions	42%	16%	58%	16%	74%	20%	0%

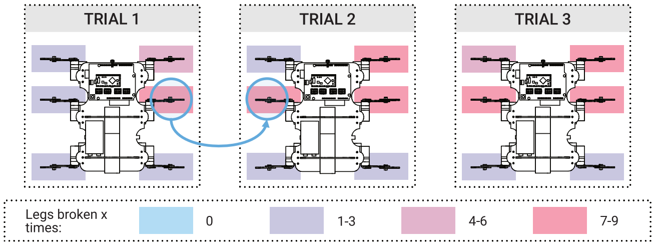

Figure 6. Legs at the front and center of the robot were preferred. Participants mostly started with the right middle leg (towards their body), followed by the leg on the opposite side. This approach tries to maintain the functionality.

Participants' Punishment Behavior

Participants listed both reasons for and against punishment as a teaching technique. Notably, positive emotions were only mentioned for dazzling and scolding. For example, P2 said that dazzling was *"somehow cool"* and *"fun"*. Moreover, dazzling the robot elicited feelings of having power from two participants, e.g. P42 claimed: *"When being dazzled, he resisted but then I felt a little powerful"*. Participants also mentioned reasons for using dazzling as punishment technique due to its feasibility and abstractness. For example, two participants preferred dazzling over the other techniques because it *"does not destroy him and he can still walk"* (P36) and *"you have a sense for yourself and [...] it's an humane punishment"* (P43). Two participants stressed that dazzling *"inflicts more abstract pain than you are used to because you don't really do this with humans"* (P17), and requires to only *"press a button"* (P26).

On the other hand, the majority of participants also signified concerns with respect to punishment. These concerns can be divided into economic, emotional, and social reasons. Concerning economic reasons, 11 participants showed inhibitions to break the legs since they wanted to avoid destroying it.

19 participants expressed emotional reasons. On the one hand, ten participants experienced *"awkward"* (P17) feelings during scolding the robot. For example, P25 thought that *"you can scold a dog or maybe someone else but, but a robot, I found that difficult [...] and I couldn't think of anything [to say]"*. Only one participant stated awkward emotions for dazzling and none for mutilation. On the other hand, none of the participants found scolding *"discomforting"*. Two participants felt uncomfortable dazzling the robot and seven to mutilate it. For example, P25 found it *"really bad. I mean [...] it's a thing.*

But somehow, that's infringing". 12 participants expressed sympathy for the robot. For example, P2 found dazzling was *"awful because it wriggled all the time"*. P3 *"felt a bit sorry for the robot"*. Seven participants regarded the punishment as cruel. Two participants thought that dazzling had *"something of torture"* (P25). In contrast, seven participants criticized the mutilation technique since it was *"too brutal"* (P42).

Finally, participants indicated social reasons against punishment. Two participants wanted to avoid making a bad impression on the experimenter, for example P2 *"did not want to come over as a sadist"*. Two others felt that the robot's reaction conveyed that it was wrong to punish it. For example P36 explained: *"Well, if it was just a piece of metal, I would just have broken off a bit. But he kind of made facial expressions."*

Applicability of Punishment for Teaching Robots

In the last part of the interview, participants were asked whether they could imagine punishment as a teaching technique for robots. Seven participants explicitly stated that they did not have any problems with punishing robots. P26 found dazzling a suitable teaching technique because it *"doesn't cause irreversible damage"* and is *"generally applicable"*. Seven participants could imagine applying punishment but named actual learning success as a prerequisite.

However, participants also expressed doubts in using punishment. First and foremost, seven participants would not use mutilation since it *"obstructs the robot"* (P3). Five participants expressed concerns that punishing robots might also affect the way humans treat other humans since punishing could *"encourage behavioral patterns so that people are also scolded when they commit mistakes"*. P25 considered punishment to be negatively connoted since *"a person who carries out a punishment isn't completely in control of himself"*. 11 participants also regarded other teaching techniques as more promising. For example, P42 suggested that *"the robot should be motivated to learn using positive stimuli"*. P26 pointed out that more sensitive people may suffer from disadvantages because they are more reluctant to use punishment. Two participants were afraid that robots might eventually fight back when being treated badly.

LIMITATIONS

Examining participants' attitude towards robot punishment requires that participants actually experience this punishment. To justify the robot punishment, we framed participants with a robot learning story in a lab setting. Since the robot's ability to walk is affected by the punishment, not all participants may have accepted the learning task as a reasonable scenario for the punishment. People's behavior might differ in the field when they are indeed annoyed by a robot's behavior. Future work could investigate these scenarios, for example for vacuum cleaner robots which did not clean sufficiently. Nonetheless, as a first exploration of people's attitude towards robot punishment, our results indicate that our framing was successful in eliciting different responses towards robot punishment.

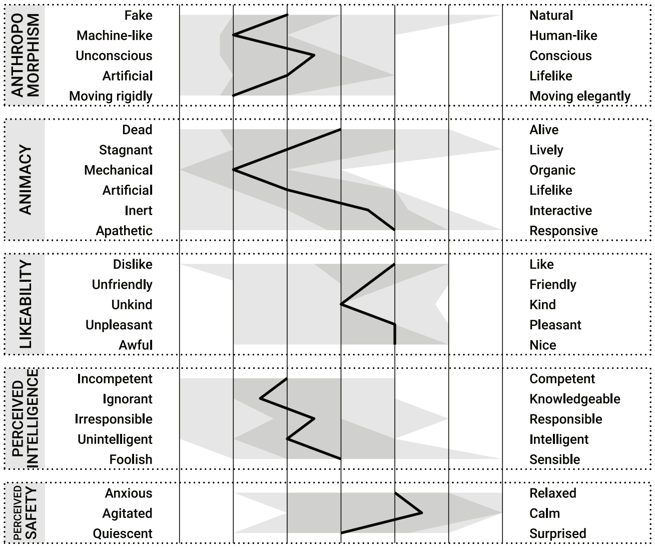

Figure 7. The Godspeed questionnaire allows to compare different robot designs regarding Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety. Our Design was rated low on Anthropomorphism and Animacy yet considered relatively responsive.

Moreover, participants' desire to conform with social expectations may have had an influence on participants' behavior and reported attitude. We counteracted these possible influences by using a pre-formulated experiment script so that the experimenter used the same responses for all participants. The experimenter was also instructed to stay in the background during the experiment.

We only informed participants after the experiment that the robot could be easily repaired. Due to the used materials (PCBs and screws), it is likely that the majority of participants expected that the robot was not destroyed irrevocably. This information is likely to have affected participants' perception of the punishment. In this paper, we argue that mutilation holds a higher meaning for participants because the physical change is permanent and limited. Hence, future work should investigate if participants perceive the punishment differently (1) with or without an information about the repairability of the robot or (2) based on different robot designs that make repairability more or less obvious.

In addition, the order of punishment may have influenced participants' willingness to punish the robot. That is, participants may have become more accustomed to performing punishment over the course of the experiment. Since breaking the robot's legs irreversibly changes the robot's performance, we decided not to counterbalance the punishment techniques in favor of incremental punishment. We also see the chosen order as a "logical" escalation pattern.

Participants' demographic characteristics were largely homogeneous. Future work should therefore expand the sample. In particular, the attitudes of older and less technology-savvy users should be examined. Since the participants were reluctant to destroy the robot due to its economic value, it may also be interesting to compare the punishment behavior of users with different social and economic backgrounds. As the

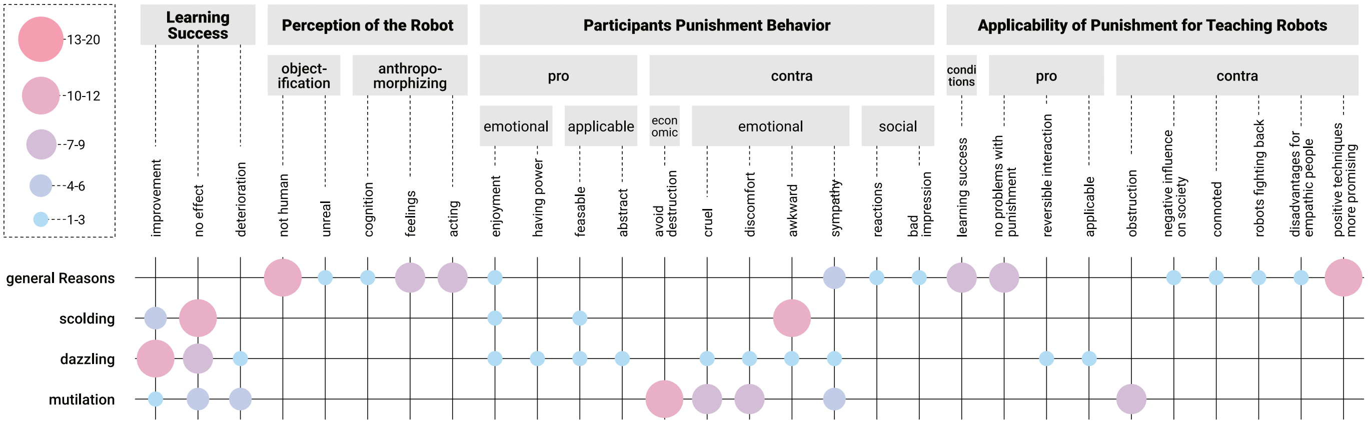

Figure 8. This Figure gives an overview of the codes derived from the interviews with the participants. The columns represent the individual codes, which are further combined to form clusters. The punishment techniques are listed in the rows. The number of participants who mentioned a certain code for a certain punishment in their interview are presented at the respective intersection by colored circles.

participants examined here are probably the first adopters of such technologies, the results still offer a relevant first insight.

While the robot physically reacts to light and touches by trembling, it does not give any feedback to being scolded. We chose this design to adapt the robot's reaction to the increasing punishment, i.e. a stronger punishment results in a stronger response. During the interviews, participants pointed out that the robot's wriggling reaction to being dazzled caused an emotional response. Hence, future work should examine the influence of the robot's reaction on participants' willingness to punish the robot. For example, a strong trembling when being scolded or no reaction after a broken leg could have a mediating effect on people's perception of punishment.

DISCUSSION

In this paper we focused on two research questions: (1) What is the boundary to using punishment as a teaching technique in HRI? (2) What are reasons for and against the usage? Literature shows that users are willing to scold, to electrify, and destroy robots, but also points out that the effects of non-harming techniques are questioned and therefore easily applied. We proposed and implemented an experience prototype, which addresses this gap. While the incremental punishment does not "kill" the robot completely (vs. destroying), the breaking a leg metaphor is more meaningful and understandable in its consequences in contrast to electric shocks.

No Clear Boundary but Clear Discomfort

When humans punish animals or other humans, they are usually more reluctant to apply corporal punishment techniques. While scolding is a common technique to educate children or dogs, uncomfortable stimuli are disapproved for human-human interaction and used only scarcely for animal education. Mutilation is, of course, clearly objected. However, concerning HRI, our findings show that humans seem to perceive different boundaries for punishing robots.

During *scolding*, participants reported to feel awkward and showed signs of uncertainty and irritation. Contrary to our expectations, *dazzling* provoked the weakest emotional response, was attributed the highest learning success, and was actively named as the preferred option during the interviews. *Mutilation* triggered more discomfort, questions, as well as nervous laughter, and was clearly disapproved by several participants.

Somewhere Between Alive and Lifeless

During the interviews, it became apparent that participants conceptualized the robot somewhere between alive and lifeless. This confirms previous findings regarding the Media Equation but also the claim by Bartneck et al. [4] that robot abuse is an exception to the Media Equation. Our results provide deeper insights into participants' reasoning to better understand the limitations of the Media Equation. Participants seemed to have difficulties to reconcile their rational knowledge (the robot is a lifeless machine) and their emotional response towards punishing the robot. Consequently, participants often provided an explanation *"but it is only a machine"* after sharing an emotional response, such as feeling sorry for the robot. This behavior could be attributed to the fact that having emotions towards lifeless objects is unintuitive or strange for most humans [39]. The results of the Godspeed questionnaire also indicate that although participants described the robot rather machine-like and mechanical, it is attributed higher likeability.

Abstract Punishment Preferred

For a general implementation of punishment, the three techniques do not differ from the robot's point of view: The robot receives an external signal, which has to be interpreted as negative feedback. Yet, our findings indicate that the human user interprets many differences in the three techniques, being primed by human-human interaction.

Scolding

While scolding elicited many signs of awkwardness among the participants based on our video analysis, the interviews suggest that most of the participants had no ethical concerns. Since verbal scolding requires high intelligence to be understood correctly, we see uncertainty during scolding as a sign for **social** non-compliance. People hesitated to conceptualize the robot as human-like and therefore were reluctant to apply human-like interaction. However, with increasingly sophisticated voice assistants entering everyday user life, scolding may become an accepted technique.

Dazzling

Dazzling seemed to be the most accepted punishment technique among participants. Participants pointed to the abstractness of the interaction, making it less comparable to human punishments. As the punishment was applied by pushing a button (a typical machine interaction), participants found it easier to execute. Since this punishment technique does not result in permanent damage, participants did not have any economic concerns. However, some participants felt uncomfortable with dazzling and mainly named the robot's reaction, a wriggling, as a reason. Again, the robot's reaction triggers a more alive impression, reinforcing participants' emotional response.

Mutilation

Bartneck et al. [4] suspected that either the perceived value of a robot or considering it "sort of alive" are responsible for participants' hesitation to destroy a robot. Our results show that mutilation elicited both very strong emotional and economic objections. Although participants described the robot as machine-like, it was perceived by 14 out 20 participants to be alive enough to provoke emotional responses such as pity and empathy. These clear emotional responses are particularly surprising since Bartneck et al. [10] found out that higher perceived intelligence and agreeableness increase humans' reluctance. According to the Godspeed questionnaire, our participants attributed the robot a lower intelligence and a medium to slightly higher likability (subdimension of agreeableness). Hence, our findings indicate that even few cues are sufficient to trigger emotional responses.

On the other hand, a majority of participants pointed out that they hesitated to destroy the robot not because of its animacy, but because they perceived it as a human-made artifact that represents time, effort, and money invested. However, participants' reactions differed strongly here. Some participants clearly expressed emotional and ethical concerns while others stated that they did not have any problems with executing punishment. Still, 14 out of 20 participants named at least one emotional response concerning mutilating the robot during the course of the interview. Again, we assume that some participants had difficulties to reconcile their emotions and considering the robot a lifeless entity.

HOW SHOULD HRI BE DESIGNED IN FUTURE?

Based on our findings, two major questions arose which challenge current design paradigms.

Could Pain-like Responses Prevent Robot Abuse?

While the literature found that service robots and personal assistants are currently often abused, our study has shown that users can connect to such machines quickly and in an empathetic manner. Reactions imitating pain clearly impacted the participants' *emotions*. Hence, feedback after a punishment is important to understand if the action actually influenced the robot. This leads to the questions whether human-like feedback can - carefully considered - prevent undesirable interaction with robots. Would children bully a service robot the same way if it cries or shows signs of fear?

Should Technology Be Designed and Treated Humanly?

This leads to the question whether we should consider punishment as an interaction paradigm at all. Even though punishment and pain are meaningful metaphors to users, they are not desirable or morally correct as design strategies. This is clearly reflected in participants' responses, who urged for more positive and helpful interactions. However, it can be argued that if technology is further anthropomorphized, both the good and bad aspects of inter-human interaction will inevitably apply to HRI. Thus, if users can thank a smart assistant for its good advice, others will scold it for bad services. If users can reward a vacuum cleaning robot for a good job, others will kick it if they are annoyed. So is it really a good idea to treat future technology humanly or should we look for design approaches which clearly differentiate between humans and machines? Do we need a paradigm shift away from "intuitive", "natural", "human-like" interaction towards more than human-centered design?

CONCLUSION AND FUTURE WORK

We presented the first implementation of a robot which allows for gradual destructive punishment motivated by the common phenomenon of abusive behavior towards robots.

While our findings show that most participants were willing to punish the robot, participants rejected the use of abusive and destructive punishment in general. Scolding, which requires participants to interact with the robot in a human-like fashion, caused discomfort, whereas mutilation was rejected for social, economic, and emotional reasons. Instead, participants preferred more abstract yet comprehensible techniques, such as the use of an unpleasant stimulus. Based on our results, we raised the question whether the intentional design of emotional responses, e.g. mimicry of pain, could prevent robot abuse.

Continuing this idea, intelligent systems as entities in future societies could create the need for emotional compensation. If intelligent systems are involved in fatal accidents, who is to blame for? Could punishment as an act of revenge trigger an emotional response that increases the perception of justice? Would we consider such an act under certain conditions? Or do we need a paradigm shift away from humanized technologies?

Future work needs to investigate the ethical implications of these paradigms for future societies. The authors' main desire is to spark a debate about a responsible and cautious approach to HRI design, taking into account the difference between human beings and machines, which should be reflected in *more than human-centered design*.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120. DOI: http://dx.doi.org/10.1609/aimag.v35i4.2513

[2] Anja Austermann and Seiji Yamada. 2008. “Good robot”, “bad robot” – Analyzing users’ feedback in a human-robot teaching task. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, New York, NY, USA, 41–46. DOI: http://dx.doi.org/10.1109/ROMAN.2008.4600641

[3] James Auger. 2014. Living with Robots: A Speculative Design Approach. *Journal of Human-Robot Interaction* 3, Article 1 (Feb. 2014), 23 pages. DOI: http://dx.doi.org/10.5898/JHRI.3.1.Auger

[4] Christoph Bartneck and Jun Hu. 2008. Exploring the abuse of robots. *Interaction Studies* 9, 3 (2008), 415–433. DOI:http://dx.doi.org/10.1075/is.9.3.04bar

[5] Christoph Bartneck, Takayuki Kanda Kanda, Omar Mubin, and Abdullah Al Mahmud. 2007. The perception of animacy and intelligence based on a robot’s embodiment. In *2007 7th IEEE-RAS International Conference on Humanoid Robots*. IEEE, New York, NY, USA, 300–305. DOI: http://dx.doi.org/10.1109/ICHR.2007.4813884

[6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. DOI: http://dx.doi.org/10.1007/s12369-008-0001-3

[7] Christoph Bartneck, Juliane Reichenbach, and Julie Carpenter. 2006. Use of Praise and Punishment in Human-Robot Collaborative Teams. In *RO-MAN 2006 The 15th IEEE International Symposium on Robot and Human Interactive Communication*, K. Dautenhahn (Ed.). IEEE, New York, NY, USA, 177–182. DOI: http://dx.doi.org/10.1109/ROMAN.2006.314414

[8] Christoph Bartneck, Juliane Reichenbach, and Julie Carpenter. 2008. The carrot and the stick: The role of praise and punishment in human–robot interaction. *Interaction Studies* 9, 2 (2008), 179–203. DOI: http://dx.doi.org/10.1257/000282803322157142

[9] Christoph Bartneck, Chioke Rosalia, Rutger Menges, and Inèz Deckers. 2005. Robot abuse – a limitation of the media equation. Proceedings of the Interact 2005 Workshop on Agent Abuse, Rome. (2005).

[10] Christoph Bartneck, Michel Van Der Hoek, Omar Mubin, and Abdullah Al Mahmud. 2007a. "Daisy, daisy, give me your answer do!" – Switching off a robot. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, New York, NY, USA, 217–222. DOI: http://dx.doi.org/10.1145/1228716.1228746

[11] Christoph Bartneck, Marcel Verbunt, Omar Mubin, and Abdullah Al Mahmud. 2007b. To kill a mockingbird robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. ACM, New York, NY, USA, 81–87. DOI: http://dx.doi.org/10.1145/1228716.1228728

[12] Cynthia Breazeal. 2001. Affective interaction between humans and robots. In *European Conference on Artificial Life*. Springer, Berlin, Heidelberg, Germany, 582–591. DOI: http://dx.doi.org/10.1007/3-540-44811-X_66

[13] Cynthia Breazeal. 2002. Regulation and entrainment in human-robot interaction. *The International Journal of Robotics Research* 21, 10-11 (2002), 883–902. DOI: http://dx.doi.org/10.1177/0278364902021010096

[14] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175. DOI: http://dx.doi.org/10.1016/S0921-8890(02)00373-1

[15] Cynthia Breazeal. 2004. Social interactions in HRI: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34, 2 (May 2004), 181–186. DOI: http://dx.doi.org/10.1109/TSMCC.2004.826268

[16] Drazen Brscić, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Takayuki Kanda. 2015. Escaping from children’s abuse of social robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI ’15)*. ACM, New York, NY, USA, 59–66. DOI: http://dx.doi.org/10.1145/2696454.2696468

[17] Maya Cakmak, Crystal Chao, and Andrea L. Thomaz. 2010. Designing Interactions for Robot Active Learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (June 2010), 108–118. DOI: http://dx.doi.org/10.1109/TAMD.2010.2051030

[18] John S. Carton. 1996. The differential effects of tangible rewards and praise on intrinsic motivation: A comparison of cognitive evaluation theory and operant theory. *The Behavior Analyst* 19, 2 (1996), 237–255. DOI:http://dx.doi.org/10.1007/BF03393167

[19] John P. Charlton. 2009. The determinants and expression of computer-related anger. *Computers in Human Behavior* 25, 6 (2009), 1213–1221. DOI: http://dx.doi.org/10.1016/j.chb.2009.07.001

[20] Hyojin Chin and Mun Yong Yi. 2019. Should an Agent Be Ignoring It?: A Study of Verbal Abuse Types and Conversational Agents’ Response Styles. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA ’19)*. ACM, New York, NY, USA, Article LBW2422, 6 pages. DOI: http://dx.doi.org/10.1145/3290607.3312826

[21] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46. DOI: http://dx.doi.org/10.1177/001316446002000104

[22] Francisco Cruz, Sven Magg, Cornelius Weber, and Stefan Wermter. 2016. Training agents with interactive reinforcement learning and contextual affordances. *IEEE Transactions on Cognitive and Developmental Systems* 8, 4 (2016), 271–284. DOI: http://dx.doi.org/10.1109/TCDS.2016.2543839

[23] Christian Daniel, Malte Viering, Jan Metz, Oliver Kroemer, and Jan Peters. 2014. Active Reward Learning. In *Proceedings of Robotics: Science and Systems X*. Berkeley, USA. DOI: http://dx.doi.org/10.15607/RSS.2014.X.031

[24] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37. DOI: http://dx.doi.org/10.1145/3185517

[25] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. 2017. Understanding Shoulder Surfing in the Wild: Stories from Users and Observers. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4254 – 4265. DOI: http://dx.doi.org/10.1145/3025453.3025636

[26] Julia Fink. 2012. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics*. Springer, Berlin, Heidelberg, Germany, 199–208. DOI: http://dx.doi.org/10.1007/978-3-642-34103-8_20

[27] Michael A. Goodrich, Alan C. Schultz, and others. 2008. Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction* 1, 3 (2008), 203–275. DOI:http://dx.doi.org/10.1561/1100000005

[28] Matthias Hoenen, Katrin T. Lübke, and Bettina M. Pause. 2016. Non-anthropomorphic robots as social entities on a neurophysiological level. *Computers in Human Behavior* 57 (2016), 182 – 186. DOI: http://dx.doi.org/10.1016/j.chb.2015.12.034

[29] Aike C. Horstmann, Nikolai Bock, Eva Linhuber, Jessica M. Szczuka, Carolin Straßmann, and Nicole C Krämer. 2018. Do a robot's social skills and its objection discourage interactants from switching the robot off? *PloS one* 13, 7 (2018), e0201581. DOI: http://dx.doi.org/10.1371/journal.pone.0201581

[30] Peter H. Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Nathan G. Freier, Rachel L. Severson, Brian T. Gill, Jolina H. Ruckert, and Solace Shen. 2012. "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology* 48, 2 (2012), 303. DOI: http://dx.doi.org/10.1037/a0027033

[31] Irene Lopatovska and Harriet Williams. 2018. Personification of the Amazon Alexa: BFF or a Mindless Companion. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 265–268. DOI: http://dx.doi.org/10.1145/3176349.3176868

[32] Qinggang Meng, Ibrahim Tholley, and Paul WH Chung. 2014. Robots learn to dance through interaction with humans. *Neural Computing and Applications* 24, 1 (2014), 117–124. DOI: http://dx.doi.org/10.1007/s00521-013-1504-x

[33] Stanley Milgram. 1963. Behavioral study of obedience. *The Journal of Abnormal and Social Psychology* 67, 4 (1963), 371. DOI:http://dx.doi.org/10.1037/h0040525

[34] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56, 1 (2000), 81–103. DOI: http://dx.doi.org/10.1111/0022-4537.00153

[35] Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2016. Why do children abuse robots? *Interaction Studies* 17, 3 (2016), 347–369. DOI: http://dx.doi.org/10.1075/is.17.3.02nom

[36] 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, Cambridge, UK.

[37] Julia Ringler and Holger Reckter. 2012. DESU 100: about the temptation to destroy a robot. In *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction (TEI'12)*. ACM, New York, NY, USA, 151–152. DOI: http://dx.doi.org/10.1145/2148131.2148164

[38] Chioke Rosalia, Rutger Menges, Inèz Deckers, and Christoph Bartneck. 2005. Cruelty towards robots. Robot Workshop – Designing Robot Applications for Everyday Use, Göteborg. (2005).

[39] Astrid M. Rosenthal-Von Der Pütten, Frank P. Schulte, Sabrina C. Eimler, Sabrina Sobieraj, Laura Hoffmann, Stefan Maderwald, Matthias Brand, and Nicole C. Krämer. 2014. Investigations on empathy towards humans and robots using fMRI. *Computers in Human Behavior* 33 (2014), 201–212. DOI: http://dx.doi.org/10.1016/j.chb.2014.01.004

[40] Pericle Salvini, Gaetano Ciaravella, Wonpil Yu, Gabriele Ferri, Alessandro Manzi, Barbara Mazzolai, Cecilia Laschi, Sang-Rok Oh, and Paolo Dario. 2010. How safe are service robots in urban environments? Bullying a robot. In *19th International Symposium in Robot and Human Interactive Communication*. IEEE, New York, NY, USA, 1–7. DOI: http://dx.doi.org/10.1109/ROMAN.2010.5654677

[41] Suleman Shahid, Emiel Krahmer, and Marc Swerts. 2014. Child–robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend? *Computers in Human Behavior* 40 (2014), 86–100. DOI: http://dx.doi.org/10.1016/j.chb.2014.07.043

[42] David Harris Smith and Frauke Zeller. 2017. The Death and Lives of hitchBOT: The Design and Implementation of a Hitchhiking Robot. *Leonardo* 50, 1 (2017), 77–78. DOI:http://dx.doi.org/10.1162/LEON_a_01354

[43] Yutaka Suzuki, Lisa Galli, Ayaka Ikeda, Shoji Itakura, and Michiteru Kitazaki. 2015. Measuring empathy for human and robot hand pain using electroencephalography. *Nature: Scientific Reports* 5 (2015), 9. DOI:http://dx.doi.org/10.1038/srep15924

[44] Xiang Zhi Tan, Marynel Vázquez, Elizabeth J. Carter, Cecilia G. Morales, and Aaron Steinfeld. 2018. Inducing Bystander Interventions During Robot Abuse with Social Mechanisms. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. ACM, New York, NY, USA, 169–177. DOI: http://dx.doi.org/10.1145/3171221.3171247

[45] Andrea L. Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737. DOI: http://dx.doi.org/10.1016/j.artint.2007.09.009

[46] Andrea L. Thomaz, Guy Hoffman, and Cynthia Breazeal. 2005. Real-time interactive reinforcement learning for robots. AAAI 2005 workshop on human comprehensible machine learning. (2005).

[47] Blay Whitby. 2008. Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers* 20, 3 (02 2008), 326–333. DOI:http://dx.doi.org/10.1016/j.intcom.2008.02.002

[48] David Whitney, Eric Rosen, James MacGlashan, Lawson L. S. Wong, and Stefanie Tellex. 2017. Reducing errors in object-fetching interactions through social feedback. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, New York, NY, USA, 1006–1013. DOI: http://dx.doi.org/10.1109/ICRA.2017.7989121

[49] Sarah Woods, Kerstin Dautenhahn, and Joerg Schulz. 2004. The design space of robots: Investigating children's views. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*. IEEE, New York, NY, USA, 47–52. DOI: http://dx.doi.org/10.1109/ROMAN.2004.1374728

[50] Ya-Huei Wu, Christine Fassert, and Anne-Sophie Rigaud. 2012. Designing robots for the elderly: appearance issue and beyond. *Archives of Gerontology and Geriatrics* 54, 1 (2012), 121–126. DOI: http://dx.doi.org/10.1016/j.archger.2011.02.003

[51] Yuto Yamaji, Taisuke Miyake, Yuta Yoshiike, P. Ravindra De Silva, and Michio Okada. 2010. STB: Human-dependent Sociable Trash Box. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, New York, NY, USA, 197–198. DOI:http://dx.doi.org/10.1109/HRI.2010.5453196