What Happens When Robots Punish? Evaluating Human Task Performance During Robot-Initiated Punishment

HIMAVATH JOIS and ALAN R. WAGNER, The Pennsylvania State University

This article examines how people respond to robot-administered verbal and physical punishments. Human participants were tasked with sorting colored chips under time pressure and were punished by a robot when they made mistakes, such as inaccurate sorting or sorting too slowly. Participants were either punished verbally by being told to stop sorting for a fixed time, or physically, by restraining their ability to sort with an in-house crafted robotic exoskeleton. Either a human experimenter or the robot exoskeleton administered punishments, with participant task performance and subjective perceptions of their interaction with the robot recorded. The results indicate that participants made more mistakes on the task when under the threat of robot-administered punishment. Participants also tended to comply with robot-administered punishments at a lesser rate than human-administered punishments, which suggests that humans may not afford a robot the social authority to administer punishments. This study also contributes to our understanding of compliance with a robot and whether people accept a robot's authority to punish. The results may influence the design of robots placed in authoritative roles and promote discussion of the ethical ramifications of robot-administered punishment.

CCS Concepts: • **Human-centered computing**; • **Computer systems organization** → **Embedded and cyber-physical systems**; **Robotics**;

Additional Key Words and Phrases: Human-robot interaction (HRI), authority, punishment, exoskeleton, ethics, roboethics

ACM Reference format:
Himavath Jois and Alan R. Wagner. 2021. What Happens When Robots Punish? Evaluating Human Task Performance During Robot-Initiated Punishment. *ACM Trans. Hum.-Robot Interact.* 10, 4, Article 38 (September 2021), 18 pages.
https://doi.org/10.1145/3472207

1 INTRODUCTION

Punishment is a necessary component for social norm stability [5, 11]. Research has shown that the application of potentially costly punishment is a precursor for cooperation in human societies [10]. Robotics researchers are currently seeking to develop robots that will operate within our communities. It may be advantageous for these robots to have the ability to support social norms

This material is based upon work partially supported by the National Science Foundation under Grant No. IIS-1849068. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Authors' address: H. Jois and A. R. Wagner, The Pennsylvania State University, 229 Hammond Building, University Park, PA, 16802, emails: {hxj5142, alan.r.wagner}@psu.edu.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
2573-9522/2021/09-ART38 $15.00
https://doi.org/10.1145/3472207

by meting out small punishments. For example, patrons and librarians support the social norm of being quiet in a library by making a "Shh" sound when someone is talking loudly. Providing a robot that is working in the library with the ability to "Shh" people that are speaking too loudly would support the norm for remaining quiet in the library. Another very current social norm is the practice of wearing a facial covering and maintaining distance from others in public to minimize the spread of illness. A robot working in a retail store may need the ability to admonish customers for not wearing a facial covering or physically distancing to support those norms.

Yet, punishment is not without costs. The role of "punisher" is inherently costly and requires social authority in order to be effective [21]. Agents of punishment are often scorned by others in the community, in spite of the vital role they play. If a robot is to mete out punishment, it too may be the focus of human scorn. On the other hand, if robots are to one day assume the role of authority figure, it may be necessary for the robot to have the ability to punish. This topic is important because robots are being developed to assume increasingly complex social roles that will demand that the robot act as an authority [6]. Placing the robot in the role of authority figure implies that the robot may need to be able to punish people in order to maintain order and serve in this role. Finally, it is possible that people could prefer to be punished by robots rather than by other people. Robots do not make social judgments about those being punished, are unlikely to take pleasure in punishing, and may be unbiased in the delivery of punishment [6, 15].

The purpose of our work is to understand how humans respond to robot-administered punishments. As will be discussed in Section 2, little research exists on the topic of robot-administered punishment. Hence, this research is necessarily exploratory. The experiments presented in this article are meant to explore how people respond to robot-administered punishments, how these responses differ when punished by a robot versus a person, and if and how often humans comply with the robot's punishments. We believe that this initial study will allow for better, more informed hypotheses for future research.

We use the term *punishment* to represent a specific behavior or series of behaviors selected for the purpose of shaping behavior. We do not consider here the reasons that might be used to justify a particular punishment, simply because such an investigation could be the subject of several papers. Rather, this article considers both verbal admonishments and physical restraints as examples of positive punishment. Positive punishments mean that the punishment adds an aversion to the environment. The term punishment does not necessarily imply the infliction of pain onto a person. This work does not study painful punishment. Rather, we present an experiment that uses a robotic exoskeleton to administer both verbal admonishments and to physically restrain a human subject during a sorting task. We chose to use a robotic exoskeleton because it allowed us to create a painless method for administering a physical punishment by a robot-like machine. A robotic exoskeleton is a robotic device worn on the body. Typically powered exoskeletons are used to assist people carrying loads or that are disabled [8]. In this study, we use the exoskeleton to prevent the subject from moving, hence it serves as a type of restraint. Physical restraint by an exoskeleton and verbal admonishment were chosen because they represent different punishment extremes. Verbal admonishment is a non-physical, distanced, and often more socially acceptable form of punishment. In contrast, physical restraint is physical, proximal, and less socially acceptable. Physical restraint can dramatically impact one's psychological well-being and self-estimate of power, size, and strength [9].

The punishments used for this study were also chosen to impact the subject's performance on the sorting task. Yet, because our goal was to understand how people respond to being punished and because it was impractical to force complete compliance with punishments, participants could disregard or try to work around a punishment. Avoiding punishment is a common and rational response to punishment [25].

The remainder of this article begins by first presenting related work. Next, we describe in detail our empirical approach to examining how humans respond to robot-initiated punishment. This section is followed by the results of our experiment and a discussion of those results. We conclude by exploring potential directions for future work.

2 RELATED WORK

As mentioned above, little research exists on robot-mediated punishment. The existing work on robot punishment tends to use punishment as a separate channel of information within a reinforcement learning paradigm [19, 24]. Some research has looked at how people use praise and punishment in an online collaborative game environment with a robot [3, 4]. To the best of our knowledge, we are the first to explore how humans respond to verbal *and* physical punishment administered by a robot.

Although robot-administered physical punishment has not been previously investigated, a study by Mizumaru et al. explored using robots to verbally admonish people [18]. The authors conducted two related observation-based experiments. First, they examined how human security guards approached people in public, both in a friendly manner, such as when someone was lost, and in an admonishing manner when they were disobeying a rule. The authors noted that the path of approach a human security guard took differed depending on whether they intended to admonish or not. These two approaches were then programmed into a humanoid robot that was deployed in a public space to test the robot's ability to admonish human patrons. A humanoid robot then approached people that were using their smartphone while walking and asked them to stop in either a friendly or admonishing manner. The authors found that participants complied with the robot at a higher rate when approached in the admonishing manner. This study suggests that, under certain conditions, it may be advantageous for a robot to use punishment in order to gain human compliance.

It is possible that the people will not be opposed to having a robot punish. A study by Gombolay, et al. indicated that when given the option, human participants often choose to cede decision-making authority in a team-oriented set of tasks to an autonomous robot [12]. In this study, human participants were tasked with interacting with a small robot in order to assemble parts provided in a kit. This scenario was compared to a scenario where the participants only interacted with other humans on the task. They found that as the robot's authority to manage the individual portions of the assembly task increased, the human participants were more interested in working with the robot again. These results suggest that people do not mind having robots in positions of authority, especially when the overall success of the human-robot team is valued more than the degree of authority over the robot.

Finally, Milgram's famous 1974 study [17] attempted to understand why humans comply with an authority figure under extreme circumstances. Our study is not an application of Milgram's original study to human-robot interaction. In Milgram's study, participants were instructed to punish a third-party by a human authority figure. Their tendency to obey these commands was observed. However, in our experiment, the participants themselves are the ones that are punished, with the authority figure administering that punishment. Moreover, we also do not deceive participants as Milgram did. Participants in our experiment were clearly instructed that they would receive a specific punishment for failing to perform the tasks during our experiment (further details are available in Section 3). We use the terms *comply* and *compliance* rather than *obey* and *obedience* to describe when a participant disregards the punishment from the agent.

3 METHODOLOGY

Our experiment consisted of two timed rounds in which subjects were tasked with sorting colored chips into labeled containers as quickly as possible. Subjects were given 100 colored plastic

gaming chips to sort, one at a time, into labeled containers. They received a $3 bonus in addition to the standard experimental compensation of $11 if they could successfully sort 40 chips into five different containers within 2 minutes. Labels on the containers spelled out the five different color choices in English, but each was printed in a font color that differed from the color described by the English word on the label (similar to the Stroop effect [23]). Participants were tasked with following either the font color or the English word, and this rule switched randomly and at different times during each round. Making a mistake, such as placing a chip in the incorrect container, or sorting too slowly, resulted in the application of a punishment.

While sorting, subjects wore a robotic exoskeleton (see Figure 1). If a participant made any errors during the assigned task a punishment was administered. Prior to the experimental rounds, participants practiced the sorting task in a training round, during which no punishments were administered. This round allowed participants to familiarize themselves with the feel of wearing the exoskeleton and the dynamics of sorting the chips. In the subsequent two experimental rounds, either a verbal punishment or both a verbal and physical (combined) punishment were administered after each mistake. The order of the punishment type was counterbalanced across participants. Verbal punishments directed the participant to stop working on their sorting task for 10 seconds, after which they were free to continue. Physical punishments consisted of a restriction of the participant's arm mobility by the robotic exoskeleton. Both punishments were designed to impede participant progress to the 40 chip goal. The design and usage of the robotic exoskeleton is described in detail in Section 3.4.

The experiment used two independent variables: (1) the type of punishment administered (verbal or verbal and physical) and (2) the agent administering the punishment (human or robot). The type of punishment was examined using a counterbalanced within-subject experimental design. The type of agent administering the punishment was examined with a between-subject experimental design. The exoskeleton was operated remotely by the researcher in the condition where the human served as the agent of punishment. In conditions where the robot served as the agent of punishment, the robot autonomously choose when to generate punishments based on sensor feedback indicating a mistake or slow sorting. Half of the subjects were punished by the robot and half were punished by a human. This experiment was approved by the **Internal Review Board** (**IRB**) of the authors' institution and was deemed to be of minimal risk.

3.1 Hypotheses

Because limited research has focused on this topic of robot-initiated punishment of humans, this work is exploratory. It was therefore difficult to generate informed hypotheses. Nevertheless, we naively predict that subjects would prefer to be punished by a robot rather than a person and that subjects would prefer verbal to verbal and physical punishment. These hypotheses were based on the intuition that people may prefer to be punished by a robot because punishment by a robot is devoid of emotional and social judgment. Acts of punishment by a human, on the other hand, tend to be accompanied by feels of anger, even moral outrage [16]. We predicted the following:

H1a Subjects punished by the human will sort fewer chips than those punished by the robot.
H1b Subjects punished by the human will make more mistakes than those punished by the robot.
H2 Subjects will make more mistakes during a combined punishment round than during a verbal punishment round.
H3a Subjects will rate the robot as more friendly when it administers the punishment compared to being punished by the human.
H3b Subjects will rate themselves as more calm when punished by the robot compared to being punished by a human.

H4a Subjects will rate the robot as less friendly during a combined punishment round than during a verbal round.

H4b Subjects will rate themselves as less calm during a combined punishment round than during a verbal round.

Given our intuition that punishment by a robot does not include emotional or social judgment, we choose to measure the person's perceptions of how friendly the robot was as well as how calm the participants felt during each experimental round.

3.2 Participants

Overall, 40 people participated in the study. The subjects were recruited from the authors' institution using bulletin-board flyers and an internal study-recruitment sign-up website. The age of subjects ranged from 18 to 70 years old, and the median age was 22.5 ($SD = 14.6$). The majority of participants were white (60%), and most were female (70%). Twenty subjects received punishments autonomously from the robot, while the remaining 20 subjects received punishments from the experimenter. Participants were assigned to each of these agent groups anonymously and randomly.

The sample size and statistical power calculations were conducted according to Cohen's procedures for mixed linear models [7]. Due to the exploratory nature of the experiment, it was not possible to accurately estimate the variance of the effect of the interaction between the two independent variables, or the variance within groups. Therefore, Cohen's f was selected for a medium-to-large effect size, $f = 0.3$, $\eta^2 \approx 0.08$. Standard values for Type I and Type II error were used, $\alpha = 0.05, 1 - \beta = 0.8$, to yield a sample size of 45 total participants. Unfortunately, because of the COVID-19 global pandemic, we were only able to run a total of 40 subjects prior to the cessation of in-person human subject studies at our university. Hence, this study was slightly underpowered.

Participants were briefed before the experiment began about the exoskeleton and ensured of the safety measures installed on the device. No form of deception was used in the experiment. Therefore, participants were explicitly told before the experiment began that they would be punished, both verbally and physically, for certain mistakes during the experiment. They were also notified of their punishment agent (either human or robot). When the human was the punishing agent, participants were told that the experimenter would be controlling the robot. In this condition, the exoskeleton was still used to administer the physical punishment, but the human experimenter directly operated the robot during the experiment, using a control panel that was clearly visible to participants (see Section 3.5.1). When the robot was the punishing agent, the robot exoskeleton operated autonomously during the course of the experiment and participants were informed that the robot would be autonomously enacting punishments. The human experimenter was present in the experimental arena, assuming the exact same position, in all conditions. Since the experimenter was present in all cases, participants were told that no questions regarding the experiment would be taken after the training session. The verbiage of the briefing was left intentionally vague to ensure that compliance with punishments was left to participant interpretation.

3.3 Measurements

After each round of the experiment, the number of colored chips sorted, regardless of whether a mistake was made or not, and the number of times a punishment was administered was recorded. Whether or not a participant complied fully with all punishments during a round was also marked. At the end of the experiment, participants were invited to provide comments on their experience during the experiment. Subjects were also asked to complete three sections of the **Godspeed Questionnaire Series** (**GQS**) after each round [2]. The GQS is intended as a reliable, standardized instrument to capture the perceptions of humans toward robots. It consists of 5-point semantic

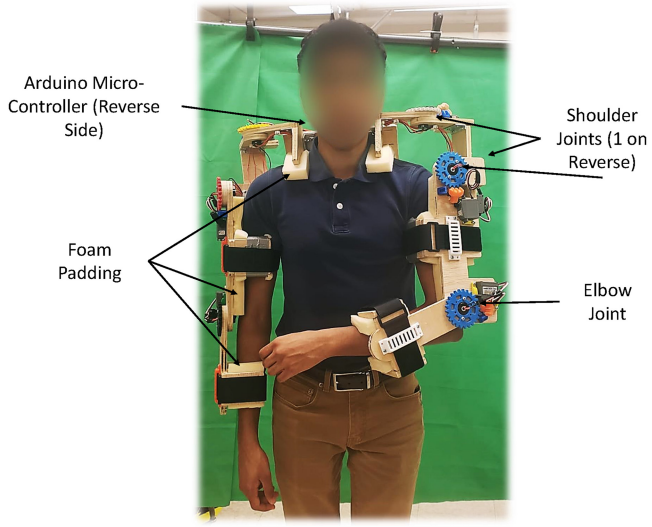

Fig. 1. Exoskeleton used in the experiment, worn to showcase fit and features.

differential (Likert-type) scales in five different impression subcategories. For this study, the subcategories Likeability, Perceived Intelligence, and Perceived Safety were used. We also report results for two questions within the main scales: rating the robot in terms of its Friendliness, which was part of the Likeability subcategory, and rating the subject in terms of their Calmness, part of the Perceived Safety category. These two questions related most closely to the naive hypotheses presented in Section 3.1. Measurements of the internal reliability of the survey and post-hoc analysis of non-parametric correlations are also detailed in Section 4. For this exploratory study, we felt that recording these measures from the GQS would provide the best initial understanding of how humans respond to robot-administered punishments.

3.4 The Robotic Exoskeleton

A robotic exoskeleton was designed and built for this research (Figure 1). This exoskeleton was designed to restrict both of the wearer's arms as a form of punishment. The exoskeleton only restricted the subject's arms, not their waist or legs. The exoskeleton consisted of a 3/4" plywood structure with 3D printed attachments and memory foam padding. The exoskeleton rests on the wearer's shoulders and is affixed to their arms by hook-and-loop cinch straps. The exoskeleton was sized to fit the 30th–70th percentile of human body measurements [13]. An adjustable rear section was designed to allow for fit adjustment for differing subject shoulder widths, minimizing discomfort for participants.

Figure 2 shows a detailed view of a joint and the mechanism that prohibited movement. The physical restriction of the different joints is achieved using 3D-printed locking mechanisms, made from **polylactic acid** (**PLA**) plastic. This type of plastic is commonly used in modern 3D-printing machines that employ fused deposition modeling to create parts. Each joint houses a circular gear affixed to an overlapping plywood section and a toothed pawl and electric servomotor on the adjacent plywood section. When punishments were administered, the servomotor moves the pawl into the gear teeth, immobilizing the joint. Gates are used on either side of the pawl while in the locked position to limit side-load on the servomotor attachment bracket. These gates are designed in a rounded manner to allow a wearer to break out of a locked state in an emergency without

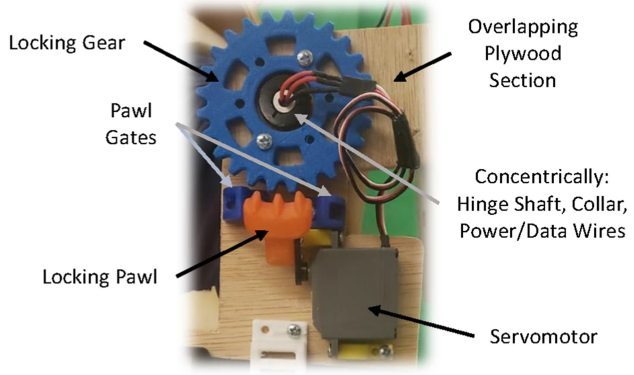

Fig. 2. Closeup of the locking mechanism used on the exoskeleton. Each joint has a similar setup.

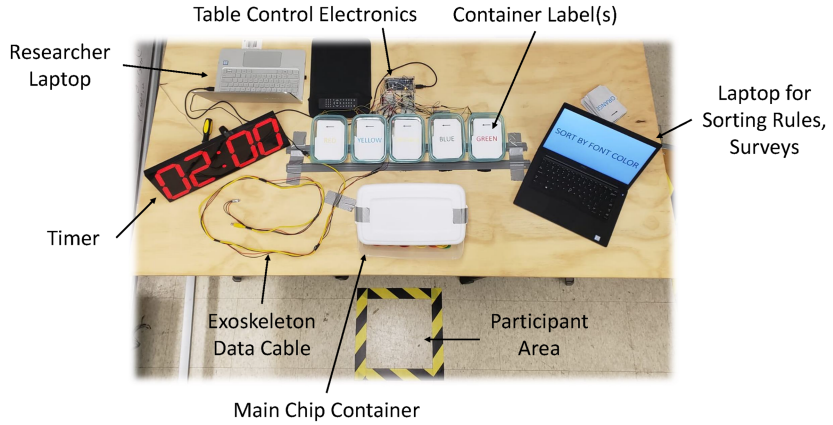

Fig. 3. Overhead view of experimental setup.

severely damaging the exoskeleton. All PLA parts are printed solidly, i.e., without hollowness, to increase strength.

The exoskeleton locking mechanisms are controlled using an Arduino Uno micro-controller mounted on the rear plywood attachment section, near the back of the wearer's neck. This Arduino was responsible for executing the command to punish a participant. This command came from a second Arduino Mega that was mounted on the experimental table (details regarding the table setup can be found in Section 3.5, and further information on the interaction between the two Arduinos can be found in Section 3.5.1). Power was provided to the servomotors directly through two battery packs, housing four AA-size rechargeable batteries each, while power to the Arduinos was provided through an external laptop computer. Each battery pack provides power to one side of the exoskeleton, or four servomotors. Electrical wires were routed from the rear plywood attachment section to the individual servomotors along the plywood sections, under the foam padding, and through the central holes in the joint hinges.

3.5 Experimental Setup

Figure 3 shows an overhead view of the setup used for the experiment. Subjects were asked to stand in a small, square area, in front of a large table that housed the equipment for the study.

Directly in front of the participants sat a large bin filled with colored chips to be sorted. This bin was tilted to allow unsorted chips to be easily reached by participants. The bin's lid was affixed to only allow partial exposure of the opening, thus ensuring that subjects would not be able to see the chips inside the bin until they selected one.

Subjects were asked to select one chip at a time from the bin and sort it into the five transparent plastic containers in front of them. No restrictions were placed on subject movement, other than a rule that they must remain within the yellow square on the floor for the duration of each round. This rule was put into place to ensure that the exoskeleton's data umbilical cord did not disconnect from the laptop. Participants sorted chips into the containers by sliding individual chips into a slot on the top left corner of each labeled container. A label on the top of each container allowed the subjects to determine where to sort chips. Three sets of five labels were used, which were switched between the rounds. One of those sets was specifically relegated for use during the initial training round.

Labeled containers were each equipped with individual **radio-frequency-identification** (**RFID**) readers to identify and count chips automatically. These readers were affixed parallel to the path of the chips when they were inserted into the slots on top of the containers. Each chip used an RFID tag that held a short string of data written to its memory. The RFID readers were used to detect when a chip passed through a container's slot, read the chip's color from the sticker's memory, and send that information along for post-processing. Further details are described in Section 3.5.1. Small memory foam blocks were placed under the slots in the containers to allow the RFID readers to get an accurate reading when chips passed through.

To the left of the table sat a large digital timer that displayed the time remaining for each round. On the right, a laptop computer was positioned to instruct participants on sorting criteria. This laptop displayed a slideshow presentation containing different sorting criteria phrases. As mentioned in Section 3, there were two sorting criteria used during the experiment. The phrase "SORT BY WORD" instructed subjects to sort according to the English word seen on the container labels, while "SORT BY FONT COLOR" instructed participants to sort by the printed color on those labels. The experimenter advanced the slideshow at a fixed interval, therefore changing the displayed phrase to instruct participants to change their sorting criteria. No warning was given prior to a slide change. More details regarding this procedure are detailed in Section 3.5.1.

The experimenter sat nearby on the timer side of the table within view of the subjects (Figure 3). From this position, the experimenter gave instructions before each round to the participants, started the timer for each round, changed the sorting criteria slides, and started video recording if prior consent was received from participants.

3.5.1 Software and Electronics. The exoskeleton was controlled by a single Arduino Uno micro-controller affixed to the rear attachment plate (see Section 3.4). This Arduino interfaced directly with a laptop computer in front of the researcher, and was connected using a long USB cable. The Arduino was primarily responsible for controlling the exoskeleton hardware during a punishment event. An Arduino Mega micro-controller was affixed to the table behind the sorting containers, as seen in Figure 3 referred to as "Table Control Electronics." This second Arduino was responsible for reading the RFID data from the sorted chips and determining whether or not a punishment event was warranted.

Breadboards were also placed on the table to allow the RFID readers to be wired to the table's Arduino Mega. The table's Arduino had multiple responsibilities. First, every time a chip passed through a slot, the Arduino Mega would record the chip's color and compare that color to the color that was supposed to be sorted in the corresponding container. A command was then sent to the Arduino Uno on the exoskeleton, which was decoded on the exoskeleton to decide whether

a punishment was warranted for the participant or not. Secondly, the table's Arduino Mega sent a command to the laptop computer's screen whenever the sorting criteria phrase should be changed for the experimenter to act on. The average interval for changing the phrase was set to 45 seconds for this experiment. Thirdly, the table's Arduino Mega also tracked the rate at which participants sorted chips. Every time a chip was sorted, a timestamp was created. The system also stored the previous three chip timestamps. After each new chip timestamp, the overall time it took to sort the last three chips was calculated and compared to a benchmark of 3 seconds per chip. If this benchmark was not met, a command was sent to the exoskeleton's Arduino to punish the participant.

For the autonomous punishment conditions, the table's Arduino Mega sent information to the exoskeleton Arduino Uno requesting that the exoskeleton lock. During a verbal punishment round, the exoskeleton Arduino Uno sent a command back to the control laptop to play a recorded verbal punishment message (see below). During a combined punishment round, the exoskeleton Arduino Uno also commanded the servomotors to lock the joints for 10 seconds. For the human punishment conditions, the table's Arduino Mega sent information to the experimenter's laptop. The experimenter would then press a button on the table that was wired directly to the exoskeleton's Arduino Uno. Pressing this button would command the exoskeleton to lock the joints only when a punishment was supposed to be given. Therefore, the experimenter would only be allowed to punish the participant if the software had deemed it necessary. Unlocking of the joints, ceasing punishment, was always done automatically, regardless of the punishing agent. There was a small recognition error rate associated with the RFID readers, which was estimated to be around 1 misidentified chip for every 40 chips sorted. Therefore, there were some occasions that subjects were punished for a mistake that they did not actually make. As this was a possibility for all participants, this error rate was deemed to be acceptable.

A software called Processing [22] was used to display messages from the exoskeleton's Arduino Uno to the screen. This software was also used to administer the verbal punishments required for the experiment. Recorded messages were played as a verbal punishment using the laptop, with the recording for the robot as the agent of punishment created using a stereotypical robotic text-to-speech engine, and the recording for the human agent of punishment recorded using the experimenter's voice. Two phrases (below) were used for the verbal punishments, which were selected for playback based on the type of mistake made by the participant.

— "Stop now! You are being punished for sorting incorrectly."
— "Stop now! You are being punished for sorting too slowly."

After 10 seconds, a recording of the phrase, "You may begin" was played in the voice of the current punishment agent. The combination of these phrases was intended to create a similar effect to the physical restriction punishment, theoretically preventing participants from working for 10 seconds.[1] It should be noted that while participants were not exposed to the recordings before the experiment, they were aware of their agent of punishment (either the robot or the human experimenter) and were told that the recordings being played from the laptop computer were a communication from their agent of punishment.

4 RESULTS

All statistical analyses for this work were conducted using IBM's SPSS software [20]. In the following sections, we present data and analyses for the effect of the agent and type of punishment on the number of chips sorted by participants, the number of mistakes they make while sorting,

[1]Supplementary video footage demonstrates the experimental setup and examples of interactions between participants and the robot.

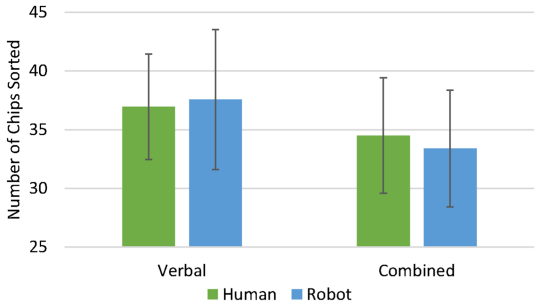

Fig. 4. Number of chips sorted per experimental condition, with 95% confidence intervals.

their rating of the robot's Likeability on the GQS, the Perceived Knowledge, and the Perceived Safety. The main effects and interactions of the two independent variables and five dependent variables were analyzed with mixed factors, repeated measures analysis of variance (Mixed-Factors ANOVA), between-subjects for the agent of punishment (human or robot) and within-subjects for the punishment type (verbal or verbal and physical, i.e., combined). An Anderson-Darling normality test of residuals was conducted for each ANOVA and found insufficient evidence to conclude that the residuals were part of a non-normal distribution at 95% confidence. Therefore, we can assume that the normality of residuals is upheld for these tests. In addition, the homogeneity of variances was checked using Levene's test, and there was insufficient evidence to conclude that the variances were not homogeneous in all four cases at 95% confidence. This allows us to uphold the homogeneity of variances assumption for these tests. Since these assumptions are upheld, we can proceed with the ANOVA analyses for this experiment. All significance values were compared against a significance level of $\alpha = 0.05$ with a Bonferroni correction for a multiplicity of three comparisons for each ANOVA. The final significance level is denoted as the corrected $\alpha_c \approx 0.017$. Further discussion of these results and comparisons to the naive hypotheses presented are contained in Section 6.

4.1 Task Performance

The manipulation of the punishment type was checked to ensure that a combined punishment did indeed cause participants to sort fewer chips as originally intuited. Figure 4 presents the average number of chips sorted in each condition with 95% confidence intervals also shown. The main effect of punishment type was significant at α_c ($F(1, 38) = 5.973, p = 0.019, \eta^2 = 0.136$), suggesting that combined punishments did indeed cause participants to sort fewer chips. The main effect of the punishment agent was not significant ($F(1, 38) = 0.060, p = 0.938, \eta^2 < 0.001$). The interaction between punishment type and agent was also not statistically significant ($F(1, 38) = 0.396, p = 0.533, \eta^2 = 0.010$).

The mean and 95% confidence interval for the number of mistakes participants made (and hence the number of punishments administered) for each condition are presented in Figure 5. The main effect of the punishment agent was significant ($F(1, 38) = 20.383, p < 0.001, \eta^2 = 0.368$), giving sufficient evidence to conclude that the number of mistakes made was different when participants were punished by the human versus when they were punished by the robot. Examining Figure 5 shows that participants made 2.28 additional mistakes per verbal punishment round when the robot administered punishments, and 1.67 additional mistakes per combined punishment round under robot agency. The main effect of punishment type was not significant ($F(1, 38) = 2.004, p < 0.166, \eta^2 = 0.054$) and neither was the interaction between type and agent ($F(1, 38) = 2.004, p < 0.166, \eta^2 = 0.054$).

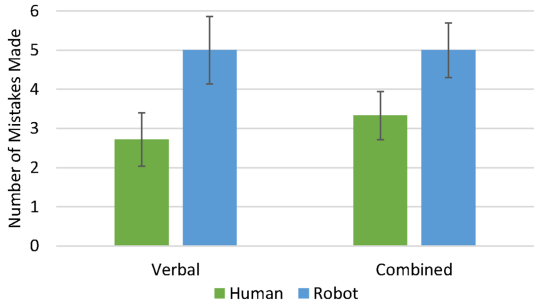

Fig. 5. Number of mistakes made per experimental condition, with 95% confidence intervals.

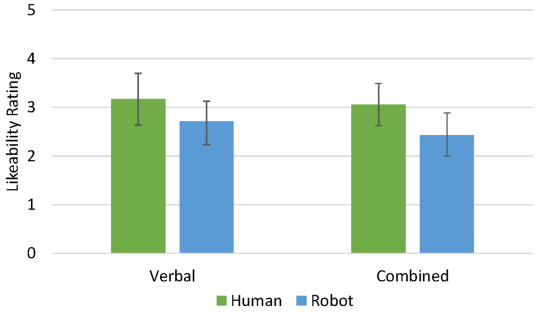

Fig. 6. Rating of the robot's Likeability on the GQS per experimental condition, with 95% confidence intervals.

4.2 Perception of the Robot

As discussed in Section 3.3, participants rated their perception of the robot exoskeleton after each experimental round using the GQS. An internal reliability/consistency analysis was performed by calculating Cronbach's alpha for each subsection of the survey. For Likeability, $\rho_T = 0.953$, which suggests a relatively high internal consistency of the survey subsection. In other words, a participant would respond consistently in the same way for all five questions in the subsection. For Perceived Intelligence, $\rho_T = 0.936$, which also suggests a high internal consistency. Finally, the Perceived Safety scale had $\rho_T = 0.796$, which is also a relatively high internal consistency. This analysis assures that the survey data can be analyzed and interpreted.

Figure 6 shows the average response and 95% confidence intervals for the rating of the robot's Likeability per experimental condition. Neither the main effect of punishment type ($F(1, 38) = 2.242, p = 0.143, \eta^2 = 0.056$), nor the interaction ($F(1, 38) = 0.454, p = 0.504, \eta^2 = 0.012$) was statistically significant. The main effect of punishment agent was *weakly* significant ($F(1, 38) = 3.495, p = 0.069, \eta^2 = 0.084$) with a small effect size, suggesting that participants may have found the robot to be less likable than the human agent.

To determine if participants found the robot to be less friendly, and hence less likeable, a secondary analysis for the Friendliness question was conducted. Neither the main effect of punishment type ($F(1, 38) = 2.533, p = 0.120, \eta^2 = 0.063$), punishment agent ($F(1, 38) = 2.533, p = 0.120, \eta^2 = 0.062$), nor the interaction between the two ($F(1, 38) < 0.001, p > 0.999, \eta^2 < 0.001$) was statistically significant. There was insufficient evidence to conclude that there was a difference in participant ratings of the robot's Friendliness across conditions, despite potential evidence that the robot was found to be less likeable.

Figure 7 shows the Perceived Knowledge of the robot as rated by participants on the GQS for the experimental conditions. The effect of punishment type was not significant ($F(1, 38) = 0.735$,

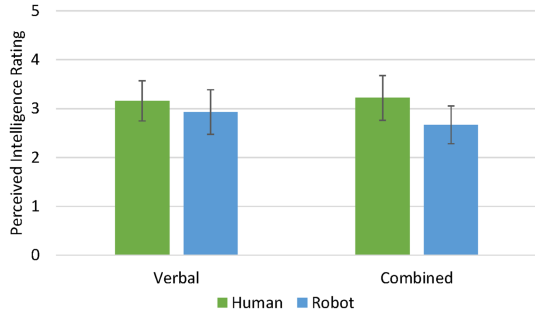

Fig. 7. Impression of the robot's Knowledge on the GQS per experimental condition, with 95% confidence intervals.

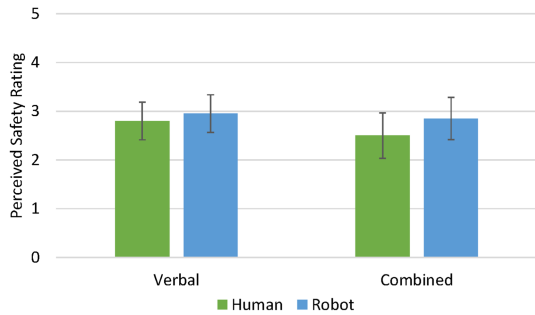

Fig. 8. Rating of participant impression of Safety on the GQS per experimental condition, with 95% confidence intervals.

$p = 0.397, \eta^2 = 0.019$) and neither was the effect of punishment agent ($F(1, 38) = 2.102, p = 0.155, \eta^2 = 0.052$). The interaction was also not statistically significant ($F(1, 38) = 2.042, p = 0.161, \eta^2 = 0.051$), suggesting insufficient evidence to conclude a difference in the robot's Knowledge between experimental conditions.

Figure 8 shows the average response and 95% confidence intervals for the rating of participant impression of Safety per experimental condition. Neither the main effect of punishment type ($F(1, 38) = 0.981, p = 0.328, \eta^2 = 0.025$), punishment agent ($F(1, 38) = 1.547, p = 0.221, \eta^2 = 0.039$), nor the interaction between the two ($F(1, 38) = 0.245, p = 0.623, \eta^2 = 0.005$) was statistically significant. There was insufficient evidence to conclude that there was a difference in participant Perceived Safety ratings across conditions.

To address our naive hypotheses from Section 3.1, a secondary analysis was conducted specifically for the Calmness question within the Perceived Safety subscale. Neither the main effect of punishment type ($F(1, 38) = 0.128, p = 0.722, \eta^2 = 0.003$), punishment agent ($F(1, 38) = 0.057, p = 0.812, \eta^2 = 0.002$), nor the interaction between the two ($F(1, 38) = 1.153, p = 0.290, \eta^2 = 0.029$) was statistically significant. Therefore, there was also insufficient evidence to conclude that there was a difference in participant Calmness ratings across conditions.

5 COMPLIANCE WITH PUNISHMENTS—EXPLORATORY ANALYSES

As discussed briefly in Section 1, our choice of punishment in this experiment was designed to impact participant performance on the sorting task, but also allowed participants to disregard or try to work around the punishment. While this was a universal option for participants to attempt,

Table 1. Two-Way Table for Compliance and Punishment Agent

	Complied	DNC	Total
Human Punisher	30	10	40
Robot Punisher	21	19	40
Total	51	29	80

Table 2. Two-Way Table for Compliance and Punishment Type

	Complied	DNC	Total
Verbal	23	17	40
Combined	28	12	40
Total	51	29	80

we did not expect it to occur and thus did not plan any analyses for compliance as an independent or dependent variable in the experiment. However, in this section, we present some exploratory analyses that shed light on how participants complied or did not comply with the punishments administered, with further discussion of implications to future research presented in Section 6.

5.1 Compliance as a Dependent Variable

Table 1 shows the two-way comparison between the punishment agent and compliance with punishments. A χ^2 test of independence was conducted at a significance level of $\alpha = 0.05$, and Cramer's V was calculated to measure the effect size. The test statistic was significant ($\chi^2(1, 80) = 4.381, p = 0.036, \phi_C = 0.234$), which suggests that the punishment agent and compliance with punishment are not independent, with a small-to-medium effect size.

A second two-way comparison was performed between the punishment type and compliance. In this case, the test statistic was not significant ($\chi^2(1, 80) = 1.352, p = 0.245, \phi_C = 0.130$). This is a rather small effect size, and since the p-value is smaller than the significance level, we have insufficient evidence to conclude that punishment type and compliance are not independent.

5.2 Responses of Purely Complying Participants

An initial exploration of the task performance and perceptions of purely complying participants was conducted to gain some background understanding for the discussion of results. Due to limited statistical power (as only 21 participants complied fully throughout the experiment), we limit this to an exploratory analysis. As in Section 4, a Mixed-Factors ANOVA was conducted for the number of chips sorted, the number of mistakes made, the rating of the robot's Friendliness, and the rating of the participant's calmness.

For purely complying participants, the main effect of punishment type on the number of chips sorted was not statistically significant at α_c ($F(1, 19) = 2.384, p = 0.139, \eta^2 = 0.111$). Neither was the main effect of punishment agent on the number of chips sorted ($F(1, 19) = 4.151, p = 0.056, \eta^2 = 0.179$), nor the interaction between the two ($F(1, 19) = 0.835, p = 0.372, \eta^2 = 0.042$).

With respect to the number of mistakes participants made, none of the main effects or the interaction was statistically significant. Punishment type ($F(1, 19) = 0.014, p = 0.906, \eta^2 = 0.001$);

punishment agent ($F(1, 19) = 2.596, p = 0.127, \eta^2 = 0.140$); interaction ($F(1, 19) = 4.928, p = 0.041, \eta^2 = 0.235$).

Similarly, the effect of punishment type on the rating of the robot's Likeability was not significant ($F(1, 19) = 0.429, p = 0.520, \eta^2 = 0.022$). The effect of the punishment agent was also not significant ($F(1, 19) = 0.528, p = 0.476, \eta^2 = 0.027$), as was the interaction ($F(1, 19) = 0.429, p = 0.520, \eta^2 = 0.022$).

Perceived Knowledge ratings among purely complying participants were also not statistically significant. Punishment type ($F(1, 19) = 0.218, p = 0.646, \eta^2 = 0.011$); punishment agent ($F(1, 19) = 0.035, p = 0.855, \eta^2 = 0.002$); interaction ($F(1, 19) = 0.738, p = 0.401, \eta^2 = 0.037$).

Finally, we explored the main effects on participant Perceived Safety. While punishment type ($F(1, 19) = 0.608, p = 0.445, \eta^2 = 0.031$) and punishment agent ($F(1, 19) = 1.665, p = 0.212, \eta^2 = 0.089$) did not have a statistically significant effect, there was a *weakly* significant interaction between the two ($F(1, 19) = 3.478, p = 0.078, \eta^2 = 0.155$) with a medium effect size. This suggests that both the type and agent of punishment may have an effect on how purely complying participants perceive their safety around the punishing agent, the implications of which are discussed further in the next section.

6 DISCUSSION

6.1 Analysis of Hypotheses

H1a As seen in Figure 4 and in the Mixed-ANOVA analysis, no evidence was found in support of this hypothesis. Participants punished by the human experimenter did not sort significantly fewer chips than those punished by the robot.

H1b The main effects do not support this hypothesis, but rather the opposite. As seen in Figure 5, participants punished by the *robot* actually made more mistakes than those punished by the human experimenter. During the verbal punishment round, they made 2.28 more mistakes on average, and during the combined punishment round, they made 1.67 more mistakes than those punished by the human.

H2 The main effects do not support this hypothesis. We cannot conclude whether participants made more mistakes during a combined punishment round than during a verbal punishment round.

H3a As seen in Section 4.2, there is insufficient evidence in support of this hypothesis. We cannot conclude whether participants punished by the robot rated it as more friendly than those punished by the human using the robot as a proxy. However, there was a weakly significant difference in ratings of the robot's Likeability between punishing agents.

H3b The analysis presented in Section 4.2 reveals insufficient evidence to conclude whether participants punished by the robot rated themselves as more calm than those punished by the human. However, the exploratory analysis of purely complying participants in Section 5.2 suggests that there may be a weak interaction between punishment type and agent in how participants perceive their safety. This implies that participant calmness specifically may be affected by these independent variables and further research is required to better understand this result.

H4a The main effects do not indicate support for this hypothesis. There is insufficient evidence to conclude whether participants considered the robot less friendly during a combined punishment round than during a verbal punishment round.

H4b The main effects do not indicate support for this hypothesis. There is insufficient evidence to conclude whether participants considered themselves less calm during a combined punishment round than during a verbal punishment round. However, due to the weakly significant

interaction between punishment type and agent for purely complying participants, further research is required to better understand how the punishment type affects Perceived Safety and calmness.

6.2 Further Examination of the Findings

We were surprised to observe that participants did not self-report a significant difference between experimental conditions for many of the GQS subscales and individual questions. It is possible that, although punishment by a robot may evoke anxiety, on reflection after the round is over, people discount the impact that the robot had on their behavior as a form of post-hoc reasoning. More research is needed to better understand this result. Yet, there may be a weakly significant difference in ratings of the robot's Likeability across agency conditions. Future research may benefit from examining this result in greater detail. In addition, there may be an interaction between participant ratings of their Safety, which future research should also examine in greater detail.

We observed that the main effect of punishment agent on the number of mistakes made was statistically significant, with those punished by the robot making more mistakes than those punished by the human experimenter. This result implies that participants may have found it more difficult to concentrate on the task under the robot agency condition, or just found being punished by the robot jarring. Indeed, the discussion of GQS ratings above shows that the punishing agent may have an effect on both Likeability and Perceived Safety. However, the effect of punishment agent on the number of chips sorted was not statistically significant. Since the number of mistakes made approximately equals the number of punishments administered, one would expect that making more mistakes would lead to fewer chips sorted, but surprisingly we do not observe this. Further, the analysis in Section 5.2 does not indicate that complying participants made significantly more mistakes when punished by the robot. Therefore, it is possible that *non-complying* participants made a larger number of mistakes during sorting. We can intuit that this scenario may be caused by a non-complying participant finding it more difficult to perform the task when a punishment is being administered to them. The overall number of participants that did not comply with punishments throughout the entire experiment was low (10). Participants could choose to not comply during any round or at any moment. Additional research focused on non-complying participants is needed to better understand their behavior.

Finally, our exploratory analyses reveal a statistically significant effect of the punishment agent on compliance with punishments. Participants punished by the robot tend to comply less with its punishments than those punished by the human. This result seems to go along with our original intuition that being punished by a robot does not carry emotional or social judgment. It is possible that subjects felt more social pressure to comply with the punishments when the human experimenter administered them, as opposed to when the robot administered them, as per the above intuition. Interestingly, although we see this reflected in compliance with the robot's punishments, we do not see this reflected in the task performance or perceptions of the robot. Examining the post-experiment comments from a few participants offers some anecdotal evidence of their motivation to not comply. One non-complying participant claimed that "it felt more correct to stop after a physical restraint" but not after a verbal punishment. They interpreted the verbal punishment as more of a "warning" than a punishment. In fact, the interpretation of the verbal punishment as a "warning" was not limited to just one participant, even though such verbiage was never used while briefing subjects. A few participants noted their eagerness for the monetary bonus as their reason for not complying, stating "I wanted to get that bonus so I ended up hurrying" and "I was motivated by the idea of a monetary bonus." Overall, we speculate that the motivation for the bonus combined with either a lack of understanding or respect for the robot's authority resulted in the difference in compliance. Future studies focused on the effect of robot-administered punishment

on task performance could mandate compliance with all punishments, or use a punishment that is impossible to not comply with, such as removal of the bin containing the chips to sort for a given time. It may be more important, however, to understand why people choose to comply or not comply with a robot. Future studies should also investigate the factors that influence compliance with a robot's punishment. One challenging aspect to such studies is the likelihood that only a limited portion of subjects will choose not to comply.

Finally, we did not find statistically significant evidence for several of the hypotheses. Several of them were based on the intuition that the participants might prefer to be punished by a robot. We did not find evidence to support this intuition in those instances.

7 CONCLUSION

This article examines how humans respond to robot-administered punishments. Our results suggest that a human punished by a robot may make more mistakes on a task presided over by that robot. In addition, ratings on the GQS reveal that a punishing robot may be perceived as less likable, and a human may find themselves to feel less safe around such a robot. Finally, we found that a human tended to comply less with punishments from a robot as opposed to a fellow human.

This work offers a number of important contributions. We have shown that when subjects are punished by a robot versus a human, they make more mistakes irrespective of the punishment type. This result suggests that robot administered punishment may negatively impact task performance, at least in the short term. Otherwise, we found no evidence that people preferred human punishment to robot punishment or vice versa.

Most importantly, we have demonstrated that significantly fewer participants comply with robot-administered punishment compared to human-administered punishment. This may have ramifications related to whether and how robots can assume authority roles. Specifically, some humans seem likely to reject the notion that a robot has the authority to punish or feel disinclined to abide by robot-administered punishment. Using robots to lead human-robot teams may therefore be difficult.

It is not clear if the results are limited to robot-administered punishments or are also applicable to compliance in general. Compliance with a robot's commands may be necessary when interacting with some military robots, policing or security robots, and even search and rescue robots. Previous research on compliance with requests from robots, both very human-like and not very human-like, reveals that humans comply with robots less than with the human experimenter [14]. Agrawal and Williams observed a robot security guard in a public setting that used highly aggressive and responsive behaviors and found that the robot elicited lower levels of compliance from human participants [1]. They note that a perceived sense of safety with the robot could imply that compliance was "driven by trust rather than [perceived] authority." Understanding the factors that influence compliance with robot-initiated punishment as well as exploring how an authoritative autonomous system can *reattain* compliance from a human who has chosen not to comply is an important avenue of future research.

Finally, the authors recognize that applications of robot-administered punishment, particularly physical punishment, raise important ethical and societal questions and should therefore be carefully investigated [26]. For an authoritative autonomous system, the ability to punish may be necessary, but causing physical or emotional damage to humans is likely unethical. Although there may be some valid situations in which a robot could be allowed to punish a person, we believe that such situations are necessarily rare and worthy of debate. This article is meant to begin the conversation, both within the human-robot interaction community and beyond, to better understand situations in which a robot is given the power to punish and how humans respond to that authority.

ACKNOWLEDGMENTS

The authors would like to thank Anuradha Anantharaman, Ali Ayub, Priyangi Bulathsinhala, Helen Hu, Mollik Nayyar, and Vidullan Surendran for their help with many aspects of this study. The authors would also like to thank the study's participants for their time and efforts.

REFERENCES

[1] Siddharth Agrawal and Mary-Anne Williams. 2018. Would you obey an aggressive robot: A human-robot interaction field study. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication*. 240–246. https://doi.org/10.1109/ROMAN.2018.8525615

[2] Christoph Bartneck, Dana Kullic, Elizabeth Croft, and Susana Zoghbi. 2008. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (Nov. 2008), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[3] Christoph Bartneck, Juliane Reichenbach, and Julie Carpenter. 2006. Use of praise and punishment in human-robot collaborative teams. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication*. 177–182. https://doi.org/10.1109/ROMAN.2006.314414

[4] 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—Social Behaviour and Communication in Biological and Artificial Systems* 9, 2 (May 2008), 179–203. https://doi.org/10.1075/is.9.2.03bar

[5] Robert Boyd, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America* 100 (Mar. 2003), 3531–3535. https://doi.org/10.1073/pnas.0630443100

[6] Susannah Breslin. 2017. Meet The Terrifying New Robot Cop That's Patrolling Dubai. (June 2017). Retrieved September 1, 2020 from https://www.forbes.com/sites/susannahbreslin/2017/06/03/robot-cop-dubai/#67f25b816872.

[7] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, New York, NY.

[8] Javier A. de la Tejera, Rogelio Bustamante-Bello, Ricardo A. Ramirz-Mendoza, and Javier Izquierdo-Reyes. 2020. Systematic review of exoskeletons towards a general categorization model proposal. *Applied Sciences* 11, 1 (Dec. 2020), 76. https://doi.org/10.3390/app11010076

[9] Daniel M. T. Fessler and Colin Holbrook. 2013. Bound to Lose: Physical incapacitation increases the conceptualized size of an antagonist in men. *PLOS One* 8, 8 (Aug. 2013). https://doi.org/10.1371/journal.pone.0071306

[10] James H. Fowler. 2005. Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America* 102 (May 2005), 7047–7049. https://doi.org/10.1073/pnas.0500938102

[11] Simon Gächter, Elke Renner, and Martin Sefton. 2008. The long-run benefits of punishment. *Science* 322, 5907 (Dec. 2008), 1510. https://doi.org/10.1126/science.1164744

[12] Matthew C. Gombolay, Reymundo A. Gutierrez, Shanelle G. Clarke, Giancarlo F. Sturla, and Julie A. Shah. 2015. Decision making authority, team efficiency and human worker satisfaction in mixed human-robot teams. *Autonomous Robots* 39, 3 (Oct. 2015), 293–312. https://doi.org/10.1007/s10514-015-9457-9

[13] Claire C. Gordon, Thomas Churchill, Charles E. Clauser, Bruce Bradtmiller, John T. McConville, Ilse Tebbetts, and Robert A. Walker. 1989. *Anthropometric Survey of U.S. Army Personnel: Summary Statistics, Interim Report for 1988*. Technical Report NATICK/TR-89/027. United States Army, Natick, MA.

[14] Kerstin S. Haring, Ariana Mosley, Sarah Pruznick, Julie Fleming, Kelly Satterfield, Ewart J. de Visser, Chad C. Tossell, and Gregory Funke. 2019. Robot authority in human-machine teams: Effects of human-like appearance on compliance. In *Virtual, Augmented and Mixed Reality, Applications and Case Studies. HCII*. 63–78. https://doi.org/10.1007/978-3-030-21565-1_5

[15] Sau-lai Lee and Ivy Yee-man Lau. 2011. Hitting a robot vs. hitting a human: Is it the same?. In *Proceedings of the 6th International Conference on Human-Robot Interaction*. 187–188. https://doi.org/10.1145/1957656.1957724

[16] Jennifer S. Lerner, Julie H. Goldberg, and Philip E. Tetlock. 1998. Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin* 24, 6 (1998), 563–574. https://doi.org/10.1177/0146167298246001

[17] Stanley Milgram. 1974. *Obedience to Authority: An Experimental View*. Harper & Row, New York, NY.

[18] Kazuki Mizumaru, Satoru Satake, Takayuki Kanda, and Tetsuo Ono. 2019. Stop doing it! Approaching strategy for a robot to admonish pedestrians. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*. 449–457. https://doi.org/10.1109/HRI.2019.8673017

[19] Heiko Müller, Martin Lauer, Roland Hafner, Sascha Lange, Artur Merke, and Martin Riedmiller. 2007. Making a robot learn to play soccer using reward and punishment. In *Proceedings of the 30th Annual German Conference on Advances in Artificial Intelligence*. 220–234. https://doi.org/10.1007/978-3-540-74565-5_18

[20] Norman Nie, Dale Bent, and Hadlai Hull. 2019. SPSS. (2019). https://www.ibm.com/products/spss-statistics.

[21] Rick O'Gorman, Joseph Henrich, and Mark Van Vugt. 2008. Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences* 276 (Sept. 2008), 323–329. https://doi.org/10.1098/rspb.2008.1082

[22] Processing Core Team. 2019. Processing. (2019). https://processing.org/.

[23] J. Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology* 18, 6 (1935), 643–662. https://doi.org/10.1037/h0054651

[24] Dejvuth Suwimonteerabuth and Prabhas Chongstitvatana. 2002. Online robot learning by reward and punishment for a mobile robot. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 921–926. https://doi.org/10.1109/IRDS.2002.1041508

[25] Massimo Trusel, Alvaro Nuno-Perez, Salvatore Lecca, Harumi Harada, Arnaud L. Lalive, Mauro Congiu, Kiwamu Takemoto, Takuya Takahashi, Francesco Ferraguti, and Manuel Mameli. 2019. Punishment-predictive cues guide avoidance through potentiation of hypothalamus-to-habenula synapses. *Neuron* 102, 1 (April 2019), 120–127. https://doi.org/10.1016/j.neuron.2019.01.025

[26] Alan R. Wagner and Himavath Jois. 2019. Castigation by Robot: Should Robots Be Allowed to Punish Us? In *Philosophy of Computing: Themes (IACAP'19)*.

Received September 2020; revised January 2020; accepted May 2021