SHOULD ROBOTS BE ALLOWED TO PUNISH US?

Himavath Jois and Alan R. Wagner
The Pennsylvania State University
University Park, PA

EXTENDED ABSTRACT

Antoine de Saint-Exupéry, a pioneering French aviator, proclaims in *The Little Prince* that, “The machine does not isolate man from the great problems of nature but plunges him more deeply into them” [1]. Autonomous robots are currently being developed to assist in teaching, provide aspects of healthcare, and work alongside soldiers in the battlefield. As these systems become more capable and ubiquitous, we must begin to decide what types of behavior will be out-of-bounds for artificially intelligent autonomous systems. For example, should an autonomous robot operating in the classroom have the ability to admonish a student that is not on task? Should an exoskeleton, a mechanical device one wears to support movement, monitor a wearer’s movements and chastise them if they are putting themselves at risk? Must an autonomous robot operating alongside soldiers on the battlefield report soldiers that abandon their duties? Our work focuses on the use of robotic exoskeletons. Since exoskeletons strap onto one’s body, interaction between the person and the robot is both intimate and physically driven.

To be clear, we are not focusing on the legal implications of punishing robots for potential crimes they commit; as Peter Asaro notes, “it is not clear that [punishing] them would achieve the traditional goals of punishment” [2]. Rather, our intent is to examine how people respond to being punished by a robot. Clearly, some social situations, such as teaching, demand that a robot have the ability to punish humans in order to accomplish its task. Yet, in other instances, such as using an exoskeleton, it seems that there should be a limit to how, when, and why a robot punishes a person. In spite of their critical importance to society, outside of science fiction, these questions have not been seriously studied.

It is reasonable to conclude that when robots do have the autonomous capability to assimilate into human society, those robots will assume roles of authority. We must therefore seek to understand the social and ethical ramifications of allowing a robot to punish a person. We investigate this problem not because we believe that robots should punish people, but rather, to understand the limits of how and when such a technology should be used. We intentionally consider an extreme scenario in which an exoskeleton restricts the wearer’s movement. Our work is meant to examine the ethical principle of autonomy which formally recognizes the importance of allowing people control over their thoughts and actions. Autonomy is meant to protect individual choice and freedom against control by the state or other people. Philosophically, our work examines how a machine might be used to limit a person’s autonomy and how people react to these limitations. Our intent is to alert the community to the possible uses these systems and to initiate the academic conversation surrounding the possibility of using an autonomous system to exact punishment. Asking questions that force us to consider the possible ramifications of a developing technology is a valuable and important exercise.

It is possible that the majority of humans may not feel opposed to having a robot punish them for their mistakes versus another human doing the same. A study by Gombolay, et. al., seems to indicate 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 [3]. This could indicate that people do not mind having robots in positions of authority, especially considering that in the study, the participants valued the overall success of the human-robot team more than the degree of authority they had over it. However, the human participants were not punished by the robot for any mistakes they made during the assigned set of tasks. We hypothesize that members of our society may actually favor a robot punishing them to a human authoritative figure.

Whatever may be the preferences of our society, this situation raises yet another important question: where will we draw the line with robot-initiated punishment? What types of punishment are considered unethical or immoral? Offering a prisoner an alternative form of punishment which restricts his or her movements to those deemed acceptable may be preferable to incarceration. Restricting the movements of prisoners of war may be preferred to POW camps. Yet wholesale restriction of a populace's movements in order to dissuade from general social ill or towards some social good is a frightening possibility. Until we examine these ethical boundaries, it will be difficult to predict the actual places where our society should choose to draw the lines.

We have designed an experiment in which subjects are asked to sort colored objects into labelled bins during three different rounds and are also punished for mistakes they make. Bins are labeled in a font color that differs from the label color name. During the experiment, subjects are either punished verbally or physically by a robot. Verbal punishment consists of a prerecorded admonishment in a robotic voice. Physical punishment, however, required the design and creation of a wearable robotic exoskeleton that could restrict the movement of a research subject's arms in a way that is not physically harmful to the subject. The exoskeleton straps to the subject's arms and has four joints for each arm, one for each elbow and three for each shoulder. The exoskeleton also has locking mechanisms at each joint, which immobilize the whole system in unison to restrict the subject during a "punishment" event. We have received IRB approval for this experiment.

We also compare *robot-initiated* punishment to *human-initiated* punishment. During *human-initiated* punishment the experimenter witnesses the subject's mistakes and applies either verbal or physical punishment. During robot-initiated punishment conditions, however, the robot autonomously applies punishment when a mistake occurs. Our intent is to examine how agency impacts the administration of punishment. Specifically, do people view a human punisher more or less favorably than a robot punisher? Psychological research suggests that people find the act of punishing as rewarding [4]. A question remains as to how the agency of the punisher impacts the punishment. Will people blame a robot as the source for the punishment? Or will the robot be viewed as an instrument of some other entity, in this case the experimenter? Finally, what characteristics, if any, will cause the punished to view the robot as an autonomous agent acting on its own accord?

References [abridged due to space limitations]

[1] A. de Saint-Exupéry, *The Little Prince*. U.S.; Reynal & Hitchcock, 1943.

[2] P. Asaro, *Robots and Responsibility from a Legal Perspective*. [Online]. Available: http://www.peterasaro.org/writing/asaro%20legal%20perspective.pdf. [Accessed Feb. 13, 2019].

[3] M. C. Gombolay, R. A. Gutierrez, S. G. Clarke, G. F. Sturla, & J. A. Shah, "Decision-making authority, team efficiency and human worker satisfaction in mixed human-robot teams," *Auton Robot*, vol. 39, Jul., pp. 293-312, 2015.

[4] K. Sigmund, C. Hauert, & M. A. Nowak, "Reward and punishment," *Proceedings of the National Academy of Sciences*, vol. *98,* no. 19, Sep., pp. 10757-10762, 2001.

SHORT ABSTRACT

Autonomous robots are currently being developed for tasks that may require those robots to assume a position of authority over humans. Our work examines the ethical boundaries of human-robot interaction in the context of robotic punishment of humans. We focus on the use of robotic exoskeletons for their physically driven interaction with a human wearer. We are investigating this issue to better understand our society's potential limits on robots in positions of authority. If such robots could be used to limit a person's autonomy, how will people react to these limitations? In addition, how ethical or moral will these robotic forms of punishment be perceived as? We hypothesize that people may actually favor a robot punishing them over another human. We have designed an experiment that intends to compare human reactions to human punishment, robot-initiated punishment, and human-initiated punishment (with a robot as a punishment tool), all through the use of a robotic exoskeleton. This exoskeleton, designed in-house, is capable of restricting the motion of a research subject's arms during a "punishment" event. During the experiment, subjects will be punished for making mistakes during an assigned sorting task. We have received IRB approval for this experiment, and hope our study will shed light on the philosophical implications of robots in authoritative states.