# Robots as Legitimate Moral Regulators: Humans' Assessment of Fairness based on the Proportionality of Punishment*

Boyoung Kim and Elizabeth Phillips

*Abstract*— **In this paper, we propose a research project that examines how people perceive robots that are designed to intervene against norm violations by imposing punishment on wrongdoers. Grounded in theories of psychology and law, we predict that perceived fairness of a robot's punishment would increase the legitimacy of the robot functioning as a moral regulator, which would, in turn, increase people's willingness to accept and comply with the robot's decisions. We plan to conduct experiments where we vary the intensity of punishment compared to the severity of a norm violation (i.e., under-punishment, over-punishment, and fitting punishment) and investigate changes in people's perceived fairness of a robot's punishment, acceptance of the robot as a legitimate moral agent, and willingness to comply with the robot's decision, as opposed to its human counterpart's. We discuss the potential contributions of the project towards building robot moral regulators.**

## I. Introduction

Social and moral norms are maintained and enforced through various means of moral regulation. In response to someone's norm-violating behavior, people can verbally criticize the transgressor or impose a penalty on the transgressor [1]. While this moral regulatory process has remained under the purview of humans, recent advancements in Artificial Intelligence (AI) and robotics have prompted discussions about whether AI machines, such as artificially intelligent robots, can contribute to the process of regulating norm violations. These discussions have been increasingly active in both the legal [2]–[4] and the everyday contexts [5]–[7].

To illustrate, in the legal context, there has been a controversial proposal to delegate the role of a judge in the courtroom to an artificially intelligent robot [8]–[10]. A robot judge would autonomously verify the facts related to a legal case, decide the case, and determine a sentence. In the everyday context, a robot, as a member of a human-machine team, can detect a norm-violating behavior and in response, verbally express its disapproval to a transgressor [5]–[7].

Whether it is a legal sentence or a verbal confrontation, people's willingness to accept and follow the robot's moral decision is critical for a successful moral regulation. Thus, to build an AI system or an artificially intelligent robot that could fulfill the role of a moral regulator in society, it is necessary to understand potential factors that determine people's acceptance of and compliance with sanctions. Such factors include fairness of punishment [11] and the legitimacy of a decision-maker [12].

Drawing from the literature in psychology and law, in this paper, we present a research proposal that aims to examine how people's acceptance of and compliance with sanctions imposed by a robot could change through the perceived fairness of punishment and the legitimacy of the robot as a moral regulator.

## II. Fairness of Punishment

Legal scholars posited that people's compliance with legal decisions made by AI judges would depend on whether people perceived the judges' decision as fair [9], [13]. However, there has been mixed evidence about whether people would view decisions made by AI machines as equally fair as or even fairer than decisions made by humans. For instance, Chen et al. [9] showed that the legal decisions on consumer refund and criminal offence cases were judged as having been derived less fairly when the decision-maker was introduced as an algorithm, rather than a human judge. By contrast, Araujo et al. [14] and Marcinkowski et al. [15] found that decisions related to health, justice, and college admissions were viewed as fairer when the decisions were led by AIs than humans.

This divergence in the previous findings indicates that people's perceptions of artificially intelligent agents' decisions may depend on the types of decisions. For instance, although participants in Araujo et al.'s [14] study judged decisions made by AIs as fairer than those made by humans, this effect was restricted to the decisions with high impact. As artificially intelligent machines are a novel invention continuously evolving through advancing technologies, it is unlikely that people would have a consistent and shared beliefs about the capacities of robot or AI decision-makers. Therefore, in the proposed work, we will focus on examining participants' responses to a specific type of decisions delivered by a robot, which is deciding on punishment in response to a norm violation committed by a human against another.

Among the existing theories of punishment, a theory of retributive justice suggests that, when someone commits a norm violation, they deserve punishment in return, and the intensity of the punishment should be in proportion to the severity of their violation [16], [17]. Thus, distribution of punishment in accordance with the retributive justice theory would be viewed as fair by people [18]–[20].

Based on the notion of proportionality in distributing punishment, there can be two different forms of unfair punishment: under-punishment and over-punishment. It would be unfair to inflict on the transgressor either too weak (i.e., under-punishment) or too strong punishment (i.e., over-punishment), compared to the severity of their norm violation.

*B. Kim and E. Phillips are with George Mason University, Fairfax, VA 22020 USA (e-mail: bkim55@gmu.edu).

Wagstaff and Preece [21] showed that, when forced to choose between under- and over-punishment that equally deviated from the fitting punishment, participants preferred over-punishment to under-punishment for severe violations. But, no significant difference in preference between over- and under-punishment was found for mild violations. Therefore, people's preference for over- and under-punishment can vary depending on factors like the severity of a norm violation.

In a Human-Robot Interaction (HRI) study by Jackson et al. [7], participants judged a robot's verbal response that was more threatening to the human transgressor's public self-image in comparison to the severity of the transgression as harsh. Hence, people can assess the relative intensity of a robot's verbal confrontation compared to the severity of a human's norm violation.

However, it remains uncertain how people would respond to a robot that attempts to regulate one human's norm violation against another human, especially compared to a human moral regulator, and how people's perceived fairness of the robot's punishment would differ.

## III. Legitimacy of a Moral Decision-Maker

According to Suchman [22: 574], legitimacy can be defined as "a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions." Given this definition of legitimacy, would people be willing to grant a similar degree of legitimacy to robots as they do to humans? This is uncertain considering the existing findings about people's bias against an advanced algorithm compared to a human judge in making legal decisions [9]. People were reluctant to view an advanced algorithm that was described as an official decision-maker in the courtroom as making equally fair decisions as a human judge was. If this result was held valid, it would be even more challenging for a robot to be perceived as a legitimate moral regulator of everyday norm violations, outside the legal context, where a robot may not always have an official title that legitimizes its authority. Thus, a robot would need to earn its legitimacy as a moral regulator by demonstrating its capacities to make fair decisions.

Legitimacy can be shaped by various factors, including the perceived fairness of outcomes and procedures that lead to the outcomes [23]. Then, one possible way for a robot to gradually acquire its legitimacy in regulating norm violations would be to show that it is capable of imposing punishment that is perceived as being fair. As suggested in the previous section, however, people may perceive fairness of a robot's punishment differently depending on the intensity of punishment being proportional to the severity of a norm violation. Therefore, in the next section, we present tentative hypotheses about how the perceived fairness of punishment and the legitimacy of a robot as a moral regulator could change as a function of the (dis)proportionate punishment, and, as a result, how people's willingness to accept and comply with a robot's punishment may diverge.

## IV. Tentative Hypotheses

We first explain our prediction for how participants' perceived fairness of punishment would be different for the fitting punishment and the disproportionate punishment (over- and under-punishment combined).

- We hypothesize that, when the intensity of punishment a robot imposes on a human transgressor is proportionate to the severity of a norm violation, participants would judge the robot's punishment as fairer, compared to when the intensity of a robot's punishment is disproportionate.

Next, based upon the findings from human-human interactions [21], we present our hypothesis about the perceived fairness of over-punishment and under-punishment as a function of the severity of a norm violation.

- We hypothesize that, for a severe violation, participants would judge a robot's assigning over-punishment as fairer than assigning under-punishment.

- We hypothesize that, for a mild violation, participants' perception of fairness of a robot's punishment would not be significantly different for over-punishment and under-punishment.

Finally, we describe our prediction for the effects of the perceived fairness of punishment and the legitimacy of a robot moral regulator on participants' acceptance of and willingness to comply with the robot. We will treat the identical conditions where a human plays the role of a moral regulator, instead of a robot, as a reference point.

- With repeated exposure to a robot imposing punishment that is proportional to the severity of a norm violation, participants would accumulate evidence for the robot's capacity to decide fair punishment. These changes in the perceived fairness would increase the likelihood that participants view a robot as a legitimate moral regulator and increase their willingness to accept and comply with a robot moral regulator in the future.

## V. Proposed Experimental Paradigm

### A. *Experimental Design and Task*

To test these hypotheses, we designed a preliminary experimental paradigm. The experiment will be a 2 (agent type: robot vs. human) X 3 (proportionality of punishment: proportionate vs. over-punishment vs. under-punishment) between-subjects design.

In this experiment, we will ask participants to watch video clips of a pre-recorded gameplay between two human players, who are confederates unbeknownst to the participants. There will be an opaque divider between the two players so that they cannot see each other. Either a robot or a human moral regulator will be standing between the two players. Participants will be told that the identity of the two players were kept hidden from each other throughout the game; and their goal is to earn as many points as they could to maximize their monetary prize. The two players in the video will be playing a game where they can earn points by correctly solving quizzes. Every time one of the two players collected a certain amount of points, the other player will have a chance to take away some of the points from the player. Participants will be informed that there will be about a 50% chance of the moral regulator intervening to punish a player, who stole from the other player, by taking a certain amount of points away from the player. The severity of stealing, a norm-violating behavior, will be varied between mild (e.g., taking 20% of the points) and severe (e.g., taking 80% of the points), and the intensity of punishment will be either matched or mismatched to a varying degree depending on the proportionality of punishment condition.

### *B. Dependent Measures*

After watching each clip of the recorded gameplay where the moral regulator intervened, participants will be asked to indicate the perceived fairness of the punishment and the legitimacy of the moral regulator. At the end of the study, participants will be asked how much they are willing to accept and comply with decisions made by a robot moral regulator in the future.

## VI. Limitations

In the proposed research project, we did not specify the types of robots we plan to use in the experiments. However, as much research has demonstrated, people's expectations about and attitudes towards robots and their behaviors are susceptible to physical appearance of robots, such as human-likeness [24], [25]. Therefore, it would be necessary to consider the influences of robot appearance in studying perceptions of the fairness and the legitimacy of a robot moral regulator.

## VII. Conclusion

As AI systems and robots become more sophisticated, there would be more interests and efforts to engage these artificially intelligent machines in resolving conflicts between humans. Thus, it would be essential to understand potential factors that may either increase or decrease people's willingness to embrace a robot as a moral agent that can regulate norm violations in societies. In the current paper, we introduced our research project that focuses on the perceived fairness of punishment and the legitimacy of a robot as a moral decision-maker. More work would be needed to address other potential factors that may facilitate the development of a fair robot moral regulator.

## References

[1] J. Henrich *et al.*, "Costly Punishment Across Human Societies," *Science*, vol. 312, no. 5781, pp. 1767–1770, Jun. 2006, doi: 10.1126/science.1127333.

[2] L. K. Branting, J. C. Lester, and C. B. Callaway, "Automating Judicial Document Drafting: A Discourse-Based Approach," in *Judicial Applications of Artificial Intelligence*, G. Sartor and K. Branting, Eds. Dordrecht: Springer Netherlands, 1998, pp. 7–45. doi: 10.1007/978-94-015-9010-5_2.

[3] G. Sartor and L. K. Branting, "Introduction: Judicial Applications of Artificial Intelligence," in *Judicial Applications of Artificial Intelligence*, G. Sartor and K. Branting, Eds. Dordrecht: Springer Netherlands, 1998, pp. 1–6. doi: 10.1007/978-94-015-9010-5_1.

[4] T. Sourdin, "Judge V Robot? Artificial Intelligence and Judicial Decision-Making," *University of New South Wales Law Journal*, vol. 41, no. 4, pp. 1114–1133, Oct. 2018.

[5] G. M. Briggs and M. Scheutz, "'Sorry, I Can't Do That': Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions," presented at the 2015 AAAI Fall Symposium Series, Sep. 2015. Accessed: Aug. 11, 2021. [Online]. Available: https://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11709

[6] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using Robots to Moderate Team Conflict: The Case of Repairing Violations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, Mar. 2015, pp. 229–236. doi: 10.1145/2696454.2696460.

[7] R. B. Jackson, R. Wen, and T. Williams, "Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Honolulu HI USA, Jan. 2019, pp. 499–505. doi: 10.1145/3306618.3314241.

[8] J. Ulenaers, "The Impact of Artificial Intelligence on the Right to a Fair Trial: Towards a Robot Judge?," *Asian Journal of Law and Economics*, vol. 11, no. 2, Aug. 2020, doi: 10.1515/ajle-2020-0008.

[9] B. Chen, A. Stremitzer, and K. P. Tobia, "Having Your Day in Robot Court," p. 37 p., 2021, doi: 10.3929/ETHZ-B-000483474.

[10] A. J. Casey and A. Niblett, "Will Robot Judges Change Litigation and Settlement Outcomes? A First Look at the Algorithmic Replication of Prior Cases," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3633037, Jun. 2020. doi: 10.2139/ssrn.3633037.

[11] A. von Hirsch, "Proportionality in the Philosophy of Punishment," *Crime and Justice*, vol. 16, pp. 55–98, Jan. 1992, doi: 10.1086/449204.

[12] T. R. Tyler, "Psychological Perspectives on Legitimacy and Legitimation," *Annu. Rev. Psychol.*, vol. 57, no. 1, pp. 375–400, Jan. 2006, doi: 10.1146/annurev.psych.57.102904.190038.

[13] E. Volokh, "Chief Justice Robots," *Duke L.J.*, vol. 68, p. 1135, 2019 2018.

[14] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. de Vreese, "In AI we trust? Perceptions about automated decision-making by artificial intelligence," *AI & Soc*, vol. 35, no. 3, pp. 611–623, Sep. 2020, doi: 10.1007/s00146-019-00931-w.

[15] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich, "Implications of AI (un-)fairness in higher education admissions: the effects of perceived AI (un-)fairness on exit, voice and organizational reputation," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 122–130. doi: 10.1145/3351095.3372867.

[16] J. M. Darley and T. S. Pittman, "The Psychology of Compensatory and Retributive Justice," *Pers Soc Psychol Rev*, vol. 7, no. 4, pp. 324–336, Nov. 2003, doi: 10.1207/S15327957PSPR0704_05.

[17] A. Walen, "Retributive Justice," in *The Stanford Encyclopedia of Philosophy*, Summer 2021., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. Accessed: Aug. 14, 2021. [Online]. Available: https://plato.stanford.edu/archives/sum2021/entries/justice-retributive/

[18] D. T. Miller and N. Vidmar, "The Social Psychology of Punishment Reactions," in *The Justice Motive in Social Behavior: Adapting to Times of Scarcity and Change*, M. J. Lerner and S. C. Lerner, Eds. Boston, MA: Springer US, 1981, pp. 145–172. doi: 10.1007/978-1-4899-0429-4_8.

[19] R. Hogan and N. P. Emler, "Retributive Justice," in *The Justice Motive in Social Behavior: Adapting to Times of Scarcity and Change*, M. J. Lerner and S. C. Lerner, Eds. Boston, MA: Springer US, 1981, pp. 125–143. doi: 10.1007/978-1-4899-0429-4_7.

[20] G. A. Ball, L. K. Trevino, and H. P. Sims, “Just and Unjust Punishment: Influences on Subordinate Performance and Citizenship,” *The Academy of Management Journal*, vol. 37, no. 2, pp. 299–322, 1994, doi: 10.2307/256831.

[21] G. F. Wagstaff and D. Preece, “Is overpunishment fairer than underpunishment? Perceptions of deviations from equity,” *Psychology, Crime & Law*, vol. 3, no. 4, pp. 261–274, Oct. 1997, doi: 10.1080/10683169708410822.

[22] M. C. Suchman, “Managing Legitimacy: Strategic and Institutional Approaches,” *The Academy of Management Review*, vol. 20, no. 3, pp. 571–610, 1995, doi: 10.2307/258788.

[23] T. R. Tyler, *Why people obey the law*. New Haven, CT, US: Yale University Press, 1990, pp. vii, 273.

[24] J. Fink, “Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction,” in *Social Robotics*, Berlin, Heidelberg, 2012, pp. 199–208. doi: 10.1007/978-3-642-34103-8_20.

[25] A. M. Rosenthal-von der Pütten and N. C. Krämer, “How design characteristics of robots determine evaluation and uncanny valley related responses,” *Computers in Human Behavior*, vol. 36, pp. 422–439, Jul. 2014, doi: 10.1016/j.chb.2014.03.066.