cuss how the fairness of a robot moral regulator could be improved.

Fairness of Punishment

Peoples acceptance of and compliance with an AI's decisions to sanction human transgressors would depend on whether people viewed the AI's decisions as fair (**??**). However, there has been mixed findings about whether people would view decisions made by AI machines as equally fair as or even fairer than decisions made by humans. For instance, Araujo et al. (**?**) found that decisions related to justice were perceived as fairer when the decisions were led by AIs than humans, but this effect was limited to a higher impact case (i.e., making decisions on whether a criminal lawsuit should be started) than a lower impact case (i.e., making decisions on whether a parking ticket should be given). Chen et al. (**?**) showed that decisions on consumer refund cases and criminal cases (e.g., bail, imprisonment) were judged as having been derived less fairly when the decision-maker was introduced as an algorithm, rather than a human judge. They also found, adding a hearing or increasing the transparency of how the decisions were reached could reduce but not completely eliminate the gap in the perceived fairness between an AI judge and a human judge.

In the present work, we explore an alternative approach to improve the perceived fairness of a robot moral regulator. Grounded in a theory of retributive justice (**??**), we consider that following the principle of proportionality could enhance the fairness of a robot moral regulator's decisions. The theory of retributive justice posits 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 (**??**). Punishment that follows the principle of proportionality tends to be viewed as fair by people (**???**). We can thus, better understand how people assess the fairness of a robots punishment by studying their perception of a robots (vs. a humans) punishment that either conforms or does not conform to the principle of proportionality.

Applying the principle of proportionality in distributing punishment, fair punishment would be achieved when the intensity of punishment matches the severity of a norm violation. Unfair punishment would take place in two different forms: under-punishment and over-punishment. It would be unfair to impose upon 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. A Human-Human Interaction (HHI) study (**?**) showed that, when forced to choose between under- and over-punishment that equally deviated from the fitting punishment, people's endorsement of over- and under-punishment was similar for mild violations but their endorsement of over-punishment was stronger for severe violations. Thus, the impact of over- and under-punishment on people's perception of fairness may also be influenced by various factors like the severity of a norm violation.

In a Human-Robot Interaction (HRI) study (**?**), it was demonstrated that, after a human teammate made an offensive comment to another human teammate, a robot could prevent the conflict from being aggravated by conveying a verbal rebuke to the offender. However, as the match between the severity of a norm violation and the intensity of the robot's rebuke was not the focus of the previous work, the question about the perceived fairness of a robot's response remains unanswered. In another HRI study (**?**), it was found that, when evaluating the harshness of a robots verbal response to a humans norm-violating request, participants judged the robots response that was potentially more threatening to the human transgressors public self-image compared to the severity of the transgression as harsh. These findings suggest that people can be sensitive to the relative intensity of a robots verbal confrontation compared to the severity of a humans norm violation. However, it requires further research to understand how people would evaluate the fairness of punishment delivered by a robot that addressed a norm violation caused by a human perpetrator against another human victim. Building on the previous research, we should investigate how the perceived fairness of punishment decided by a robot (as opposed to a human) influences the perceived legitimacy of the robot, and how the perceived fairness and legitimacy eventually affect peoples willingness to accept and comply with a robot moral regulator. In the next section, we introduce our working hypotheses grounded in the literature reviews we have summarized so far.

Working Hypotheses

Based on the previous findings in human-human interactions (**???**), we first offer our prediction for how participants perceived fairness of punishment would be different for the fitting punishment and the disproportionate punishment (over- and under-blaming 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 robots punishment as fairer, compared to when the intensity of a robots punishment is disproportionate.

Next, based upon the findings from human-human interactions (**?**), we present our hypotheses 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 robots assigning over-punishment as fairer than assigning under-punishment.
- We hypothesize that, for a mild violation, participants perception of the fairness of a robots punishment would not be significantly different for over-punishment and under-punishment.

Finally, we explain our prediction for the effects of the perceived fairness of punishment and the legitimacy of a robot moral regulator on participants willingness to comply with the robot.

1. With repeated exposure to a robot imposing punishment that is proportional to the severity of a norm violation, participants would accumulate evidence for the robots 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 comply with a robot moral regulator in the future.

A Conceptual Framework for Building a Robot Moral Regulator

Lastly, in this section, we introduce a preliminary framework for building a robot moral regulator that may distribute fair punishment following the principle of proportionality (**??**). As shown in Figure 1, a robot moral regulator can be programmed to assign fair punishment that matches the severity of a norm violation caused by a human perpetrator against a human victim. Once the robot imposes punishment, it could gather feedback from third-party human perceivers on whether the punishment it imposed on the perpetrator was just right, too strong, or too weak compared to the severity of the violation. Then, the robot can update the proportionality estimation system based on the feedback. This feedback loop is critical due to the dynamic nature of norms. For instance, the norm of cooperation can dynamically change over time in different groups (**?**). This implies that, when someone violates the norm of cooperation, punishment that is viewed as fitting to one group of people may not be viewed as fitting to another group. Depending on which group the transgressor belongs to, a proper and fair punishment would be different. Therefore, for a robot to be able to function as a legitimate moral regulator that successfully regulates norm violations, the robot would need to be able to flexibly adjust its proportionality estimation system.

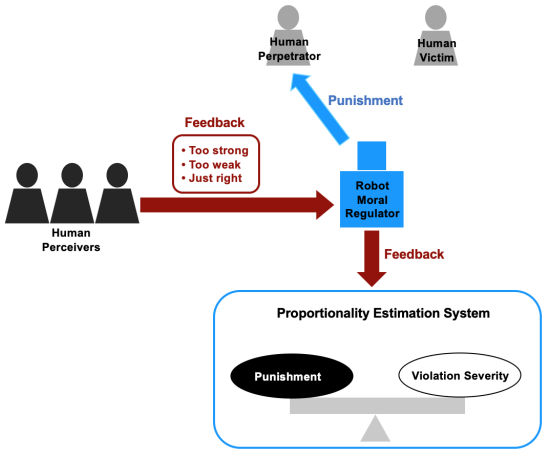

Figure 1: A schematic framework of how a robot moral regulator could update its system for generating fair punishment.

Limitations

There are several limitations that are overlooked in this paper. First, the proposed framework does not include a monitoring system that could prevent the existing human biases and errors (**??**) from being merely transferred to a robot moral regulator's decisions. The proportionality estimation system of the robot would be updated via third-party human perceivers' feedback, which may reduce the risk of decisions strictly reflecting either the victim's or the perpetrator's perspectives. However, it does not guarantee that these third-party human perceivers would be free of any biases. Second, as our discussion was focused on specific situations where victims and perpetrators of the norm-violating events are clearly determined, the proposed framework cannot explain whether and how a robot moral regulator could deal with other situations that lack such clarity.

Conclusion

As AI systems and autonomous robots become more sophisticated, there would be more discussions about whether and how these artificially intelligent machines can be properly involved in resolving conflicts between humans. Thus, it would be essential to understand potential factors that may either increase or decrease peoples willingness to embrace a robot as a moral agent that can regulate norm violations in societies. In the current paper, we suggested that the AI-HRI research community investigate the fairness and the legitimacy as the potential factors to consider in developing well-accepted artificial moral decision-makers and proposed a conceptual framework for grounding such work. Implementations of the proposed conceptual framework into autonomous robot systems would rely upon collective efforts of the experts in various disciplines of science, including Psychology, Computer Science, and Engineering.

Acknowledgment

This work was supported in part by NSF grant IIS-1909847. We thank Tom Williams at the Colorado School of Mines for his thoughtful comments on this work.

Perspiciatis iusto beatae nesciunt nisi deleniti, maiores sapiente accusantium optio id magni blanditiis fugiat sit nobis, vitae adipisci velit aspernatur omnis dolorem eligendi magnam quisquam?Ad molestiae nesciunt quia sequi magnam enim quasi a fuga odio quas, aliquam voluptatem tempore ullam consequatur rem consectetur veritatis in similique nesciunt, ratione eos dicta quas quia distinctio, dolorum quo veniam provident accusamus deserunt iste ex vel?Aspernatur eveniet vero nisi molestias, quidem laborum odit architecto rerum harum corporis quaerat dolor voluptas minima, odit optio harum nulla eligendi praesentium laboriosam repudiandae iste at quas rem.Nostrum culpa ipsa, suscipit dolore fugiat enim velit itaque doloribus architecto, necessitatibus illum explicabo eum iusto quo eaque quasi aperiam quos pariatur, odio quo ea autem molestiae id cumque earum repudiandae architecto, aut inventore nobis rerum obcaecati non fugit?Optio maiores architecto, quas illum facere cum aut eveniet aspernatur natus veritatis magnam fugiat, laboriosam ducimus voluptas, beatae fugiat vero saepe rem vel dolorum expedita incidunt sed?Ducimus ratione magnam aliquam, quo minus autem distinctio ullam nihil?Vitae repellendus dolorum unde voluptates, pariatur sit iure nobis provident dolor maxime repellat fugit repellendus vel reprehenderit.Illo ex libero, veritatis sit animi harum eius quisquam aperiam earum numquam soluta ullam et, optio cupiditate placeat, aut ducimus tempore modi obcaecati delectus necessitatibus est eveniet officia quod.