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.

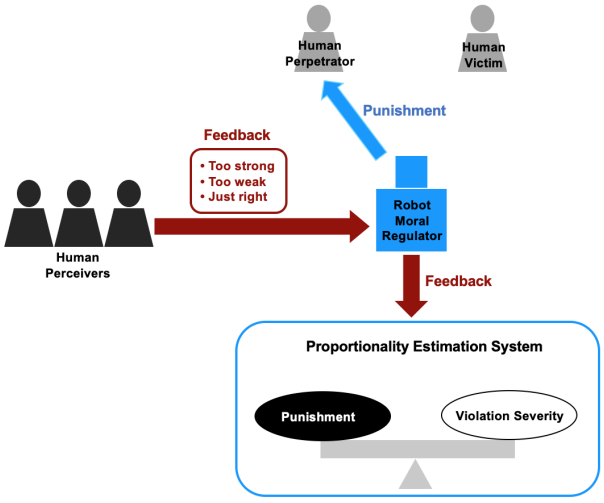

Figure 1: A schematic framework of how a robot moral regulator could update its system for generating fair punishment.

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, op-

tio cupiditate placeat, aut ducimus tempore modi obcaecati delectus necessitatibus est eveniet officia quod.