Hidden Complexities in the Computational Modeling of Proportionality for Robotic Norm Violation Response

Ruchen Wen and Tom Williams

MIRRORLab
Colorado School of Mines
Golden, CO 80401
{rwen,twilliams}@mines.edu

Abstract

Language-capable robots hold unique persuasive power over humans, and thus can help regulate people's behavior and preserve a better moral ecosystem, by rejecting unethical commands and calling out norm violations. However, miscalibrated norm violation responses (when the harshness of a response does not match the actual norm violation severity) may not only decrease the effectiveness of human-robot communication, but may also damage the rapport between humans and robots. Therefore, when robots respond to norm violations, it is crucial that they consider both the moral value of their response (by considering how much positive moral influence their response could exert) and the social value (by considering how much face threat might be imposed by their utterance). In this paper, we present a simple (naive) mathematical model of proportionality which could explain how moral and social considerations should be balanced in multi-agent norm violation response generation. But even more importantly, we use this model to start a discussion about the hidden complexity of modeling proportionality, and use this discussion to identify key research directions that must be explored in order to develop socially and morally competent language-capable robots.

Introduction and Motivation

Research has shown that language-capable robots hold unique persuasive power over humans. Robots are not only capable of influencing humans to comply with their requests and commands (Bartneck et al. 2010; Cormier et al. 2013; Rea, Geiskkovitch, and Young 2017), but can also exert moral influence over human moral norm systems (Jackson and Williams 2019, 2018; Williams, Jackson, and Lockshin 2018). It is thus important for robots to use their persuasive power in a way that is going to help regulate people's behavior and preserve a better moral ecosystem, by rejecting unethical commands (Briggs et al. 2021; Jackson and Williams 2019; Wen, Han, and Williams 2022) and calling out norm-violating behavior (Winkle et al. 2022, 2021; Jung, Martelaro, and Hinds 2015). For example, imagine a scenario where a robot and a team of humans are working together on a collaborative task, and one of the human teammate verbally abuses another team member who accidentally made a mistake. To exert positive moral influence, the robot might be expected to call out the norm violation (i.e., the insult) in order to helpfully mediate the team's dynamics.

Moreover, research has shown that robots are more persuasive when they act in a socially interactive (Fong, Nourbakhsh, and Dautenhahn 2003), sociable (Breazeal 2004) or socially agentic manner (Jackson and Williams 2021), by leveraging human-like social cues (Chidambaram, Chiang, and Mutlu 2012; Ghazali et al. 2019) and strategies (Srinivasan and Takayama 2016). This is critical when considering how robots should be leveraging their persuasive power in moral context. When rejecting immoral commands, for example, if robots only focus on explaining the morality of actions without considering the social strategies they could be using during that communication, their communication could be less effective, and they themselves could be viewed less favorably. For instance, Jackson et al. (Jackson, Wen, and Williams 2019) found that if robots give miscalibrated norm violation responses (i.e., the harshness of the responses does not match the actual norm violation severity), they will be perceived less favorably.

To appropriately calibrate the harshness of their norm violation responses, robots can use the same types of strategies that humans use, including social norms like *Politeness*. In Brown and Levinson's Politeness Theory (Brown, Levinson, and Levinson 1987), humans regularly negotiate the level of threat to one another's *Face*, which is the public image that a person wants to preserve and enhance in front of others (Brown, Levinson, and Levinson 1987). Face includes Positive Face, which is a person's wish for a desirable self-image, and Negative Face, which is a person's wish to be free from imposition and to have freedom of action (Brown, Levinson, and Levinson 1987).

From the Politeness Theory perspective, calling out a norm violation can be highly face-threatening, as it threatens both Positive Face (by making the violator looks bad in front of others) and Negative face (by appearing to intend to control the violator's behavior). Thus, it is crucial for robots to address norm violations in a way that the moral value (i.e., how much positive moral influence their response could exert) is proportional to the social value (i.e., how much face threat might be imposed by their utterance), in order to exert positive moral influence and avoid exerting unintentional negative moral influence.

Indeed, proportionality is one of the fundamental and

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

universal moral motives underlying human social-relational psychology (Rai and Fiske 2011), which is "the motive for rewards and punishments to be proportionate to merit, benefits to be calibrated to contributions, and judgments to be based on a utilitarian calculus of costs and benefits". To the best of our knowledge, however, there has been no previous work on computational modeling of how humans reason about proportionality during communication – or how robots could do the same.

In this paper, we begin to consider how such a computational model might be developed, and show the substantial complexity and nuance behind what might at first glance seem a simple problem. We present a simple (naive) mathematical model of proportionality which could explain how moral and social considerations should be balanced in multi-agent norm violation response generation. We then analyze the components of this model and procedurally explain how each is more complex than might immediately meet the eye. Through this process, we identify key research directions that must be explored in order to develop socially and morally competent language-capable robots.

A Preliminary Model of Proportionality

To begin, let us consider how we might model the utility of a candidate speech act that could be used to address a norm violation in a multi-agent context. Given our desire for proportionality, we can start by assuming that overall utility should include components for both the moral benefits of the response, and the social benefits (that is, the negation of the social harms). This model can be represented as:

$$\mathcal{U}_{\mathcal{A}} = \mathcal{U}_{\mathcal{MA}} + \mathcal{U}_{\mathcal{SA}}$$

where $\mathcal{U}_{\mathcal{A}}$ denotes the utility of a speech act $\mathcal{A}$, $\mathcal{U}_{\mathcal{MA}}$ denotes the *moral* utility of a speech act $\mathcal{A}$, and $\mathcal{U}_{\mathcal{SA}}$ denotes the *social* utility of a speech act $\mathcal{A}$.

Moral utility is positively related to moral benefits. In potentially multi-agent contexts in which robots need to address norm violations in front of zero or more human observers, the moral benefits of the robots' response come from accurately correcting misconceptions for each observer (including the violator). Thus, the calculation for moral utility can be represented as:

$$\mathcal{U}_{\mathcal{MA}} = \sum_{i=1}^{|O|}(|S_a - S_i| - |S_a - S_c|)$$

Here, $\mid O \mid$ denotes the total number of observers (including the norm violator), S_a denotes the actual norm violation severity. This could be learned from human data (Sarathy, Scheutz, and Malle 2017; Wen, Siddiqui, and Williams 2020), learned in one shot through language, or calculated through some type of utilitarian analysis. S_c denotes the norm violation severity that the speech act $\mathcal{A}$ is trying to convey. This could be estimated from empirical evidence of human perceptions or other computational models of politeness theory (Danescu-Niculescu-Mizil et al. 2013)). Finally, S_i denotes the norm violation severity that is perceived by observer O_i (the i-th observer). Each observer's prior beliefs about a norm violation's severity could be estimated from prior behaviours from that specific observer.

We might also expand this to:

$$\mathcal{U}_{\mathcal{MA}} = \sum_{i=1}^{|O|}((|S_a - S_i| - |S_a - S_c|) - \beta|S_a - S_c|).$$

Here, β denotes the weight of the dishonesty penalty, which might be affected by the robot's role. For example, dishonesty might be more greatly penalized for a teacher robot than a tour guide robot. To maximize moral utility, robots thus need to be honest with the norm violation severity: the norm violation severity they are trying to convey should be exactly (or very close to) the actual norm violation severity.

Social utility is negatively related to the amount of face threat. The more people there are to observe a face threat (and the more the violator cares about their perception by each observer) the great the social penalty they will suffer from a harsh norm violation response. Thus, the calculation for social utility can be represented as:

$$\mathcal{U}_{\mathcal{SA}} = -\sum_{i=1}^{|O|}(I(O_i, v) \times F(v, \mathcal{A})).$$

Here, $I(O_i, v)$ denotes the importance for a norm violator v to maintain a positive social image in front of the observer O_i. This could be estimated from the relational/organizational hierarchy or perceived power dynamic between the violator and the observer. $F(v, \mathcal{A})$ denotes the amount of face threat the speech act $\mathcal{A}$ imposes to the violator v. This could be estimated from empirical evidence of human perceptions of language of different types, or through computational models of politeness theory (Danescu-Niculescu-Mizil et al. 2013)).

Additional Sources of Complexity

The simple model presented in the preceding section aims to balance moral benefits against the loss of social benefits. While this model seems, on its face, to capture the basics of proportionality, it obscures a wealth of complex considerations that would need to be addressed in practice. In this section, we will discuss these hidden complexities, and speculate as to how these considerations might need to be captured in a more complex model.

Different types of observers

In the proposed model, we treat every observer equally, in terms of contributing to the total moral and social utilities. In real-life scenarios, there may be different types of observers, who may need to be treated differently from other observers. In addition to bystanders as naively considered in the preliminary model, we should also consider other types of special observers:

- The norm violator;
- People who are (or will be) negatively impacted by the norm violation, (i.e., victims); and

- People who might not be aware of the norm and have facilitated or conducted the same norm violation.

By considering the different types of observers, we can understand the ways our model might need to be adjusted accordingly. For example, it may be more important to focus on correcting the violator than to worry about correcting (likely unobserved) misconceptions that could be held by other observers. If this is the case, we might need to assign different weights on the benefits of correcting different observers.

We also may need to consider face threat to observers beyond the violator. For example, when a norm violation is called out, some observers may also feel face threaten or experience other kinds of social awkwardness if they are not aware of the norm. In this case, the total amount of face threat a speech act may cause should also include the potential face threat that the observers may receive.

Finally, victims are also important to consider when we calculate the utility of a norm violation response. For instance, calling out a norm violation, in general, may help the victims to reduce (or even avoid) the potential harm that the norm violation may cause to them. However, victims may not want somebody else to speak for them in certain cases, which might hurt both their positive and negative face by making them look incompetent and taking away their chances to defend themselves. Therefore, besides the correction benefits, the model should take all of these dimensions into account for calculating moral and social utilities.

Scaling of face threat with the number of observers

In the proposed model, the total amount of face threat is simply the summation of values over all observers. However, in reality, the amount of face threat does not tend to increase linearly with the number of observers. For example, losing face in front of 102 people and 112 people might be different, but the size of this difference is likely not as large as the difference between losing face in front of two people and losing face in front of twelve people. Given the fact that face threat likely scales nonlinearly with the number of observers, models of proportionality may need to include discount factors into their utility models.

Potential benefits of face threat

When calculating the social utility of a speech act, we consider all face threats to the violator as negative impacts and try to minimize face threat to increase utility. While this may seem reasonable at first glance, it may not always be truly beneficial to minimize face threat to a violator. First, if the violator has already caused harm, some face threat to the violator might be beneficial to everybody emotionally, for seeing the violator being called out (especially in the presence of the victims). Also, from the Confucian Ethical perspective, receiving face-threatening responses (e.g., blame-laden moral rebukes) for norm violations could help violators to cultivate their "heart of shame" (Zhu et al. 2020), which is one of the key components of Confucian moral self-cultivation. Moreover, from a pedagogical perspective, harsh responses might create a stronger impression and thus be more effective in helping people learn and grow.

While a certain amount of face threat to the norm violator could potentially be beneficial, adding it to the moral utility model is still challenging. It is unclear how much moral benefit this face threat provides; and indeed, over-weighting this benefit could have clear negative effects (e.g., robots that intentionally seek opportunities for public shaming). Models of proportionality may require careful calibration of ostensible benefits of violator-directed face threats, and how this calibration may depend on various relational and social contextual factors.

Discussion

So far we have discussed a list of possible source of complexity in the modeling of proportionality. In fact, those complexities are not only important for generating norm violation responses, but also for other type of moral communication and other aspects of moral competence, such as moral reasoning and decision making. Given the relevance to those broader topics, it is thus worth discussing the more general concerns and challenges of enabling moral competence in social robots.

In this study, we assumed that robots already had a substantial amount of prior knowledge about moral and social norms (i.e., some ground truth). However, where these norms come from and how robots should acquire this prior knowledge is an open question. Because people who have different culture backgrounds often have adhere to different moral and social norms (or do so in different ways), one potential solution is to use participatory design to collect information about which specific norms *particular communities* would follow in different situations. Such an approach would be compatible with recent calls to explore Design Justice (Costanza-Chock 2020; Ostrowski et al. 2022) and Engineering Justice (Leydens and Lucena 2017; Williams and Wen 2021) approaches to robot design. Even people from the same culture may have different moral and social beliefs due to individual life experience. In some cases, these differences might cause misunderstandings and conflicts, while in other cases, they might not negatively affect interpersonal communication. Thus, it is a challenge that robots should not only understand these differences, but more importantly, know when to seek common ground and when to preserve differences.

Additionally, in this work, our intention is to consider how we might enable robot competence to exert positive moral influence in order to help preserve and cultivate more harmonious human-robot eco-systems. While having robots calling out norm violations could be an effective approach to achieve our goals, it also bring concerns about robots "norm policing". A long term goal of this research area must be to balance the real need to call out norm violation responses without further turning robots into surveillance machines or developing tools that could be misused by law enforcement and other forms of state oppression.

Summary

In this paper, we presented a simple (naive) mathematical model of proportionality which could explain how

moral and social considerations should be balanced in multi-agent norm violation response generation. We then used this model as a starting point to consider the hidden complexity of modeling proportionality. These considerations discussed above represent key research directions that must be explored in order to develop socially and morally competent language-capable robots.

Acknowledgments

This work was funded in part by Young Investigator award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research.

References

Bartneck, C.; Bleeker, T.; Bun, J.; Fens, P.; and Riet, L. 2010. The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. *Paladyn*, 1(2): 109–115.

Breazeal, C. 2004. *Designing sociable robots*. MIT press.

Briggs, G.; Williams, T.; Jackson, R. B.; and Scheutz, M. 2021. Why and How Robots Should Say 'No'. *International Journal of Social Robotics*, 1–17.

Brown, P.; Levinson, S. C.; and Levinson, S. C. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Chidambaram, V.; Chiang, Y.-H.; and Mutlu, B. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 293–300.

Cormier, D.; Newman, G.; Nakane, M.; Young, J. E.; and Durocher, S. 2013. Would you do as a robot commands? An obedience study for human-robot interaction. In *The 1st international conference on human–agent interaction*.

Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A Computational Approach to Politeness with Application to Social Factors. In *51st Annual Meeting of the Association for Computational Linguistics*, 250–259. ACL.

Fong, T.; Nourbakhsh, I.; and Dautenhahn, K. 2003. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4): 143–166.

Ghazali, A. S.; Ham, J.; Barakova, E.; and Markopoulos, P. 2019. Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance. *Advanced Robotics*, 33(7-8): 325–337.

Jackson, R. B.; Wen, R.; and Williams, T. 2019. 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*, 499–505.

Jackson, R. B.; and Williams, T. 2018. Robot: Asker of questions and changer of norms. *Proceedings of ICRES*.

Jackson, R. B.; and Williams, T. 2019. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 401–410. IEEE.

Jackson, R. B.; and Williams, T. 2021. A theory of social agency for human-robot interaction. *Frontiers in Robotics and AI*, 267.

Jung, M. F.; Martelaro, N.; and Hinds, P. J. 2015. 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*, 229–236.

Leydens, J. A.; and Lucena, J. C. 2017. *Engineering justice: Transforming engineering education and practice*. John Wiley & Sons.

Ostrowski, A. K.; Walker, R.; Das, M.; Yang, M.; Breazea, C.; Park, H. W.; and Verma, A. 2022. Ethics, Equity, & Justice in Human-Robot Interaction: A Review and Future Directions. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 969–976. IEEE.

Rai, T. S.; and Fiske, A. P. 2011. Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, 118(1): 57.

Rea, D. J.; Geiskkovitch, D.; and Young, J. E. 2017. Wizard of Awwws: Exploring psychological impact on the researchers in social HRI experiments. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*.

Sarathy, V.; Scheutz, M.; and Malle, B. F. 2017. Learning behavioral norms in uncertain and changing contexts. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE.

Srinivasan, V.; and Takayama, L. 2016. Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 4945–4955.

Wen, R.; Han, Z.; and Williams, T. 2022. Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-based Relational Norms. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 353–362.

Wen, R.; Siddiqui, M. A.; and Williams, T. 2020. Dempster-shafer theoretic learning of indirect speech act comprehension norms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10410–10417.

Williams, T.; Jackson, R. B.; and Lockshin, J. 2018. A Bayesian Analysis of Moral Norm Malleability during Clarification Dialogues. In *CogSci*.

Williams, T.; and Wen, R. 2021. Human Capabilities as Guiding Lights for the Field of AI-HRI: Insights from Engineering Education. In *AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (AI-HRI)*.

Winkle, K.; Jackson, R. B.; Melsión, G. I.; Brščić, D.; Leite, I.; and Williams, T. 2022. Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In

Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, 120–129.

Winkle, K.; Melsión, G. I.; McMillan, D.; and Leite, I. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 29–37.

Zhu, Q.; Williams, T.; Jackson, B.; and Wen, R. 2020. Blame-laden moral rebukes and the morally competent robot: A Confucian ethical perspective. *Science and Engineering Ethics*, 26(5): 2511–2526.