Blame-Laden Moral Rebukes and the Morally Competent Robot: A Confucian Ethical Perspective

Qin Zhu (EGALITE, Humanities, Arts & Social Sciences)

Tom Williams (Computer Science)

Blake Jackson (Computer Science)

MIRRORLab
Mines Interactive Robotics Research

égalité
Ethical Governance and Learning in Technology and Engineering

Introduction

- Robots perceived as (or expected to be) moral agents
- Humans may take cues from robot teammates as to what norms apply within their shared context
- A truly socially integrated robot must be able to clearly communicate its willingness to adhere to shared moral norms including its willing to communicate its objection to others' proposed violations of such norms, through, for example, blame-laden moral rebukes.
- The ability to respond to unethical human requests using blame-laden moral rebukes as a criterion for AMA

Empirical Studies

- Empirical studies have shown that robots are able to influence, persuade, or coerce humans in different ways:
    - forego a previously desired action if a robot protests against it (Briggs and Scheutz, 2014)
    - The persuasive capability of such robots has been shown to be especially powerful for social robots (Midden and Ham, 2012)
- Our recent work has suggested that robots may be able to *unintentionally* influence the moral norms that humans believe to apply within their current context (Williams, 2018)

Our empirical work…

- Participants were asked to imagine commanding a robot to perform an action that was typically impermissible, and in their case, ambiguous ("Destroy the computer" in an environment containing two computers), and to imagine the robot responding in a way that addressed the ambiguity but not the impermissibility ("Do you mean the one on the left or the one on the right?"), thus implicitly condoning the requested action.

Our empirical work…

- Before and after reading this dialogue, participants were asked whether they thought the hypothetical robot believed such an action would be permissible and would comply with such an action, and whether they themselves believed such an action would be permissible.

Our empirical work…

- We found that not only did participants more strongly believe that the robot would believe that such an action would be permissible (and comply with it) after reading the described dialogue, but, critically, they indicated that they themselves more strongly believed the action to be permissible.

Implications

- If robots do not consider the moral implications of what is presupposed by their utterances, they may accidentally persuade their human teammates to abandon or weaken certain moral norms within their current context.
- Robots must be able to assess the permissibility of requested actions even when those requests are ambiguous.
- If robots are able to identify such morally unacceptability underlying ambiguous requests, and determine that asking for clarification would thus be problematic, how should that robot respond instead?

Empirical Studies (continued…)

- Robots that attempted repairs after perceived politeness norm violations from human teammates led to heightened awareness of those violations, which improved conflict resolution (Jung et al., 2015).
  - "Whoa, man, that was inappropriate. Let’s stay positive."
  - "Dude, what the heck! Let stay positive.”
  - Other possible responses such as rebuke: “stop that; this isn’t the place for that!” (Jehn, 1997)

How severely *should* a robot respond to a norm violation, especially violation of a moral norm?

The Relationship between Politeness and Persuasiveness

- Some researchers have found polite forms such as indirect requests to be particularly persuasive, especially with children (Kennedy et al., 2014), others have found no such relationship (Lopez et al., 2017) or even a negative relationship between politeness and persuasion, such as in healthcare contexts (Lee et al., 2017).

Complementary evidence from psychological and social sciences…

- When an issue is perceived to be of high importance, assertive direct requests are perceived as less threatening than usual, and are more effective in persuasion than more polite indirect requests (Burgoon et al., 1994), which in serious contexts can be perceived as "weak" and "too polite" (Lakoff and Ide, 2005; Tsuzuki et al., 1999).

Complementary evidence from psychological and social sciences… (continued)

- Kronrod et al. (2012) found that people are more persuaded by assertions when they already agree on the importance of an issue, and more persuaded by polite language when they are not yet convinced.

What we have learned…

- First, language-enabled robots have the power to *unintentionally* persuade based on how they respond to norm violations, in a way that may accidentally weaken humans systems of moral norm.
- Second, robots that can identify these norm violations may be able to *intentionally* wield this persuasive power in order to try to strengthen those same norms.
- Finally, when robots do so, they may need to *calibrate* the politeness or severity of their response based on how severe the norm violation is perceived to be by those the robot wishes to convince.

Blame-Laden Rebukes: A Confucian Reflection or Interpretation

The Role of Robots in the Society: A Confucian Perspective

- (Perceived) moral agents: moral self, personhood
- Technologies: social practicality

A Confucian (Role) Ethics of Robots

- The responsibilities of a person are often prescribed by the roles (e.g., father, son, engineer) assumed in specific communal contexts.
- As a true teammate, a morally competent robot has a role ethics of “caring” about the cultivation of the moral selves of other teammates.
- Social robots have a role ethics of helping human teammates to better reflect on what kind of people they are becoming and what virtues are cultivated in themselves when they make specific requests.
- A good friend has the role ethics of remonstrating with you when the friend sees you committing a wrongdoing.

The Master said, “A clever tongue and fine appearance are rarely signs of Goodness.”

CONFUCIUS. Analects: With Selections from Traditional Commentaries (Translated & Annotated) (Hackett Classics) (p. 2). Hackett Publishing. Kindle Edition.

A Virtue of Reciprocity

- Blame-laden moral rebukes may also allow human teammates to cultivate a virtue of reciprocity (*shu*, 恕).
- Different from the Christian Golden Rule, the virtue of reciprocity states ethical principles on *what not to do* assuming human teammates do not wish for others to humiliate them (Liu, 2018).
- An embodied moral psychology of Confucian shame (Seok, 2016); Mencius – “cultivating the heart of shame (*xiuwu zhixin,* 羞恶之心)”

Timely Moral Remonstrations

- For Neo-Confucianists such as Wang Fuzhi, timely moral remonstrations are crucial. When a slightly selfish desire arises, Wang suggested that that desire will recede after immediate blame.
- Without timely blame-laden moral rebukes, the “moral ecology” of the human-robot system can be negatively affected which will further develop vices rather than virtues in human teammates.
- Blame only if it helps

Blame as A Opportunity for Self-Cultivation

- However, blame-laden moral rebukes are not the only strategy robots may use to ensure adherence to moral norms, and are not, in fact, our ultimate aim.
- Confucianists distinguish the petty person (*xiaoren*, 小人) and the exemplary person (*junzi*, 君子) in terms of their distinct reactions to blame. Unlike the petty person, the exemplary person turns blame into an opportunity for self-cultivation.

From Robot-Generated Blame to Self-Blame

- From the Confucian perspective, the ultimate goal is to shift the vehicle for moral development from robot-generated blame (via blame-laden moral rebukes) to opportunities for “self-blame” (wherein humans may critically examine their own behaviors).

Thank you very much!
Let us know if you have any questions.

Qin Zhu (qzhu@mines.edu)
Tom Williams (twilliams@mines.edu)

MIRRORLab
Mines Interactive Robotics Research

égalité
Ethical Governance and Learning in Technology and Engineering