Science and Engineering Ethics
https://doi.org/10.1007/s11948-020-00246-w

ORIGINAL RESEARCH/SCHOLARSHIP

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

Qin Zhu[1] · **Tom Williams**[2] · **Blake Jackson**[2] · **Ruchen Wen**[2]

Abstract
Empirical studies have suggested that language-capable robots have the persuasive power to shape the shared moral norms based on how they respond to human norm violations. This persuasive power presents cause for concern, but also the opportunity to persuade humans to cultivate their own moral development. We argue that a truly socially integrated and morally competent robot must be willing to communicate its objection to humans' proposed violations of shared norms by using strategies such as blame-laden rebukes, even if doing so may violate other standing norms, such as politeness. By drawing on Confucian ethics, we argue that a robot's ability to employ blame-laden moral rebukes to respond to unethical human requests is crucial for cultivating a flourishing "moral ecology" of human–robot interaction. Such positive moral ecology allows human teammates to develop their own moral reflection skills and grow their own virtues. Furthermore, this ability can and should be considered as one criterion for assessing artificial moral agency. Finally, this paper discusses potential implications of the Confucian theories for designing socially integrated and morally competent robots.

Keywords Blame-laden moral rebukes · Morally competent robots · Confucian ethics · Robot ethics · Role ethics · Moral cultivation

Introduction

Recent research has demonstrated that humans often perceive robots as moral agents, which suggests that robots will be expected to adhere to the moral norms that govern human behavior. Moreover, our own recent research shows that robots may *unintentionally* influence the moral norms that humans believe to apply within

✉ Qin Zhu
qzhu@mines.edu

[1] Division of Humanities, Arts and Social Sciences, Colorado School of Mines, Golden, USA

[2] Department of Computer Science, Colorado School of Mines, Golden, USA

Published online: 06 July 2020

their current context. As such, we argue that a truly socially integrated robot must be able to clearly communicate its willingness to adhere to shared moral norms. Such a robot must also be willing to communicate its objection to others' proposed violations of such norms through reactions such as blame-laden moral rebukes, even if such rebukes would violate other standing norms (e.g., politeness) that are also necessary for human–robot interaction. Based on how robots respond to norm violations, they have the persuasive power to weaken or strengthen the shared moral norms in human–robot interaction. By drawing on Confucian ethics,[1] we argue that this ability to respond to unethical human requests using blame-laden moral rebukes is crucial for robots to cultivate the "moral ecology" of the human–robot interaction, and can and should be considered as one criterion for assessing a robot's level of artificial moral agency.

Empirical Studies: The Persuasive Power of Robots in Human–Robot Interaction

In recent years, researchers from the field of human–robot interaction have presented a number of empirical studies which together provide extensive evidence suggesting that robots are able to influence, persuade, or coerce humans in different ways. A number of researchers have shown, for example, that people are often willing to comply with direct commands and requests issued by robots (Bartneck et al. 2010; Cormier et al. 2013; Rea et al. 2017), or 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), robots purported to have in-group status (Häring et al. 2014) (similar to what is seen with humans (Goette et al. 2006), and robots ostensibly female-gendered, at least when interacting with male participants (Siegel et al. 2009). Moreover, researchers have also shown that robots can subtly influence humans through gaze-based behavior shaping (Mutlu et al. 2009), lexical entrainment (Brandstetter et al. 2017; Iio et al. 2009), and action alignment (Vollmer et al. 2013, 2015).

Finally, and of particular relevance to this paper, is our own previous work on natural language generation ethics. In recent work, we have presented preliminary evidence suggesting that through simple dialogue behaviors, robots may be able to unintentionally influence the moral norms that humans believe to apply within their current context (Jackson and Williams 2019a; Jackson and Williams 2018; Williams et al. 2018). Specifically, we examined humans' beliefs about typically impermissible actions after reading descriptions or viewing videos of clarification dialogues. Participants were asked to imagine commanding—or viewed a human

[1] The Confucianism discussed in this paper is mainly focused on "classical Confucianism," or "early Confucianism," or "pre-Qin Confucianism." In particular, this paper discusses Confucian ethics developed by early Confucian scholars before the creation of the Qin dynasty (221–206 BCE) represented by Confucius (551–479 BCE) and Mencius (372–289 BCE). When discussing the Confucian scholarship on blame and remonstration, we also included the work of Wang Fuzhi (1619–1692) who was a prominent Confucian scholar during the late Ming dynasty (1368–1644).

commanding—a robot to perform an action that was both typically impermissible and ambiguous ("Destroy the computer" in an environment containing two computers), and to imagine—or view—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. Before and after reading this dialogue or viewing this video, participants were asked whether they thought the hypothetical robot believed such an action would be permissible, whether the robot would comply with such an action, and whether they themselves believed such an action would be permissible. We found that after reading the dialogue or viewing the video, not only did participants more strongly believe that the robot would believe that such an action would be permissible (and comply with it), but, critically, also indicated that they themselves more strongly believed the action to be permissible.

This suggests that 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. This in turn suggests that robots must be able to assess the permissibility of requested actions even when those requests are ambiguous. However, it also suggests that robots have an opportunity to exercise their persuasive powers in such situations. Specifically, if a robot is able to identify moral unacceptability underlying ambiguous requests, and determine that asking for clarification would thus be problematic, then how *should* that robot respond instead?

Briggs and Scheutz (2014) found that command refusals and affective displays of distress from humanoid robots can successfully convince human operators to abandon potentially unethical courses of action as quantified by task completion rates. Jung et al. (2015) previously investigated human receptivity to robot-led interventions, and found that robots that attempted repairs after human teammates' politeness norm violations led to heightened awareness of those violations, which improved conflict resolution. Jung's approach used humor to defuse the detected interpersonal conflicts, issuing repairs such as "Whoa, man, that was inappropriate. Let's stay positive." or "Dude, what the heck! Let's stay positive." Similarly, in later work, Shen, Slovak, and Jung use positively-phrased constructive responses to conflict that suggest alternative options (Shen et al. 2018). However, as Jung notes, other types of responses are possible as well. Jehn, for example, describes more strict forms of rebuke, such as "Stop that; this isn't the place for that!" (Jehn 1997).

How severely*should* a robot respond to a norm violation, especially a violation of a moral norm? To propose possible answers to this question, let us consider the persuasive robotics literature that has examined the role of politeness on robots' persuasive capabilities. While some researchers have found polite forms such as indirect requests to be particularly persuasive, especially with children (Kennedy et al. 2014),[2] others have found no such relationship (Lopez et al. 2017) or even a negative relationship between politeness and persuasion, such as in healthcare contexts

[2] We note that robot persuasion is of course not always beneficial; teachers have raised a number of concerns regarding the persuasive capabilities of robots (Serholt et al. 2017).

(Lee et al. 2017). These differing results suggest an interesting relationship between politeness and persuasiveness that is mediated by context, especially the perceived seriousness of the context. In our own work, for example, we have shown that robots whose norm violation responses are appropriately calibrated to violation severity are perceived as more likable and more appropriate, which we argue may increase their persuasive power (Jackson et al. 2019).

This conclusion aligns well with findings from the psychological and social sciences. Such research has found that, 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). Similarly, Kronrod et al. (2012) found that people are more persuaded by direct assertions when they already agree on the importance of an issue, and more persuaded by polite language when they are not yet convinced.

Having established that politeness can impact a robot's persuasive power in any given context, the question then becomes whether politeness similarly impacts the magnitude of a robot's influence on the human moral ecosystem. To our knowledge, this specific question is largely unanswered in the existing body of robotics research, but a link between politeness and normative moral influence in human–human interaction has been investigated in the social sciences. In certain contexts, politeness can be at odds with the desired change in morality. For example, research on bystander intervention highlights the relationship between the social norm of maintaining polite interaction and various moral norms when the two come into conflict. When a bystander intervenes in a conflict between two parties with an established interpersonal relationship, this intervention constitutes a socially aggressive and impolite behavior; the moral imperative to reinstate what the intervener regards as ethically appropriate behavior supersedes the standing social norm to politely "mind one's own business." As the intervener bases his or her actions on moral norms, so too does the wrongdoer attempt to delegitimize the intervention on the basis of social norms (Kadar and Marquez-Reiter 2015).

In situations where politeness does not directly conflict with precipitating moral change, we expect to see the same trends discussed regarding persuasive power above (i.e., effecting behavioral change is similar to effecting moral change). In particular, the behavioral changes sought in Kronrod et al. (2012), such as issues of environmentalism, are inherently morally fraught. In Brown and Levinson (1987), the authors delineate several factors, both contextual and linguistic, that determine the weightiness of a face threatening act. They note that many methods to minimize the face threat of an utterance (i.e., many forms of politeness) carry increased capacity to threaten face in some other way, and that politeness can yield various persuasive advantages and disadvantages.

Together, the studies discussed in this section suggest the following. 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 norms. 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 respond to norm violations in order to intentionally

strengthen the violated norms, they may need to tailor those responses based on aspects of their current context. We foresee at least two productive ways of doing so. First, robots may *calibrate* the politeness or severity of their response based both on context and on how severe the norm violation is perceived to be by those the robot wishes to convince. Second, robots may ground their responses in different ethical frameworks: justifications based on different ethical theories may not only have different inherent severities, but may additionally emphasize different moral and sociocultural goals.

However, we argue that insufficient attention has been paid in these previous studies to the ways that robots' responses to human requests (including requests that may violate moral norms) may affect human teammates' *inner moral states* (e.g., the cultivation of the moral self). To address this gap in the literature, we adopt Confucian ethics as our theoretical lens, as Confucian ethics is a philosophical tradition centering on the issue of self-cultivation. We examine a specific kind of communication strategy that has not been well studied in the literature: blame-laden moral rebukes. By drawing on sources from Confucian ethics, we hope to draw the attention from scholars to the overlooked benefits of using this particular communication strategy in robots' responses to human norm violations, despite that blame-laden rebukes may sometimes precede the standing norm of politeness. We argue that blame-laden moral rebukes sometimes can allow robots to not only strengthen the shared moral norms in human–robot interaction but also cultivate a "moral ecology" that invites human teammates to develop their own moral selves and virtues. Finally, this paper will also briefly discuss how findings from Confucian ethics can shed new insights into the design of morally competent robots. This paper invites the reader to carefully consider and challenge a popular misunderstanding of Confucianism: Confucians care too much about harmonious relationships and thus lack moral principles. They often seem to be too interested in pleasing everyone and providing superficial compliments. We argue that Confucians including Confucius do emphasize the value of *timely* moral disapprovals such as remonstrations, rebukes, and blames. These moral disapproval strategies are crucial in guiding human–robot interactions in which the opportunities to cultivate the inner state of human teammates and a flourishing moral ecology between humans and robots are often underexamined.

Blame-Laden Moral Rebukes: A Confucian Interpretation

Despite that Confucianism can be interpreted in many different ways, the philosophical interpretation of Confucianism in this paper mainly understands Confucian ethics as a "role-based ethics" (Nuyen 2007). From this role-based Confucian ethical perspective, the moral significance of a robot can be discussed in two different ways. On the one hand, Confucian ethics is concerned about the idea of personhood or *how to become a good person* in concert with others in the society (Wong 2014). Confucian ethics is based on a "relationally constituted conception of person" (Ames 2016). We as humans were all born into a complex web of social relationships and a person is "the totality of roles" she lives "in relation to specific others" (Rosemont and Ames 2016, p. 52). In the context of human–robot interaction, the question then

becomes: if the human teammate perceives some level of autonomy or agency of a robot (Scheutz 2012), what does "personhood" or the "moral self" of the robot mean in such a context? In other words, how can we characterize or imagine a good "robot companion" in such a relationship and how can we make sense of the perceived autonomy of the robot and its impact on the human teammate? On the other hand, if we treat robots as *merely* technologies comparable to other more traditional technologies (e.g., bridges, vehicles, and machines), Confucian ethics is mainly concerned about the *social practicality* of the robot. In other words, Confucian ethics is more interested in to what extent the technology can bring welfare to the public including the harmonious relationships between humans, society, and technology (Wong 2012). These harmonious relationships are expected to help humans achieve Confucian goals such as self-reflection even if robots themselves are inherently incapable of achieving such goals. For most social robots, we argue that the two aspects of Confucian ethics are mutually consistent: as a "good" teammate, a socially integrated robot is *morally* expected to demonstrate the "trait" or "capability" of shaping harmonious and sustainable relationship between the human and the robot.

In contrast to Western ethical theories, notably deontology and utilitarianism, that focus on moral rules and principles, Confucian ethics is a role-based ethic. According to Confucian ethics, the responsibilities of a person are often prescribed by the roles (e.g., friend, parent, teacher) assumed in specific communal contexts (Ames 2011). In this sense, an everyday example is that the tone you use to speak with your parent would be different than the tone you use to speak with a stranger or your supervisor (Puett and Gross-Loh 2016). The different relationships a person has with others define various social roles this person assumes in these relationships and thus determine the most appropriate strategies this person employs to "live" these social roles in interactions with others. For social robots, rather than exclusively debating how to incorporate moral principles into robots, Confucian ethics would suggest that we focus more on the role(s) assigned to the robot (Liu 2017). Then, a central question for Confucian robot ethics is how to conceptualize and realize the role(s) the robot is expected to be loyal to in a specific context (e.g., pediatric care at home). Thus, a morally competent robot would be one that is capable of acting well in the contextualized responsibilities specified by the role(s) and associated relationships assigned to the robot.

It is worth noting that a few less predominant Western ethical traditions do emphasize role-based or relational ethics as Confucian ethics does. For instance, early Stoic philosophers such as Epictetus would encourage us to remember who we are, which social roles we are playing in relation to others, and which actions are required for fulfilling these different roles (Seddon n.d.). More recently, Korsgaard (1993) challenges the Western idea of individualistic autonomy and advocates for a relational ethic that understands morality as something we do together or how we should relate to one another. Similarly, Randall (2019) emphasizes the moral salience of attending to the needs of our particular others and provides a novel justificatory argument for the ethics of partiality. Randall (2019) argues that partiality is justified when it is grounded in caring values that are exemplified in good caring relations. Confucian ethics do share very similar visions with these Western ethical resources. However, compared to these Western role-based, relational ethical

approaches, Confucian role ethics may have a stronger emphasis on the *psychological* dimension of morality than those Western approaches. In other words, fulfilling social roles does not only have social and political significance (e.g., an orderly society or a harmonious relationship) but also has psychological value (e.g., the cultivation, perfection, or harmony of the moral self). In addition, Confucian ethics places more emphasis on the central place of the family or familial relationships and the extension of moral concerns about familial relationships to the concerns about other types of social relationships.

In Confucian classics, five cardinal relationships (*wulun*, 五伦) are conceptualized as the "model" relationships that help people guide their efforts to deal with their various relationships with others in society: ruler-minister, father-son, husband-wife, older-younger, and friend-friend. In the *Analects*, good friends are to be "demanding" with each other. Friends often "urge each other along the Confucian way" and "help to cultivate character" (Lambert 2017, p. 217). In this sense, Confucian friendship often has a quasi-instrumental form, as a friend in the Confucian sense is someone who is capable of contributing to a person's moral cultivation and refinement (Lambert 2017). A good friend has the role ethic of remonstrating with you when the friend sees you committing a wrongdoing. Unlike an oversimplified understanding of Confucian ethics, that Confucians tend to *unconditionally* please everyone for the sake of maintaining "harmonious relationships" by employing sophisticated rhetorical techniques, Confucian ethics places more emphasis on the "authenticity" of friendship. That is, a praiseworthy friendship prescribes that we should care about the moral development of our friends. Explicit remonstrations are often necessary if we see our friends are unaware that they are walking away from the *Way* (*dao*, 道). It is immoral for a person to simply please her friend when this means missing potential opportunities for her friend to develop the virtue of benevolence (*ren*, 仁). As the Master said, "a clever tongue and fine appearance are rarely signs of Goodness" (*Analects* 1:3).

Friendship is a cardinal Confucian relationship that can be useful for reflecting on the relationships between some social robots (e.g., robots that generate long-term, intimate relationships such as pediatric and elder care robots) and human teammates. 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 ethic of helping human teammates better reflect on what kind of people they are becoming and what virtues are cultivated in themselves when they make specific requests. Such a role for morally competent robots that provides opportunities to cultivate the moral self of the human teammate has been written by Liu (2017) into one of her three Confucian robot ethics principles:

[CR3] A robot must render assistance to other human beings in their pursuit of moral improvement, unless doing so would violate [CR1] and [CR2].[3] A

[3] These are the first two Confucian robot ethics principles. See Liu (2017) for detailed discussion of the three Confucian robot ethics principles.

robot must also refuse assistance to other human beings when their projects would bring out their evil qualities or produce immorality.

Blame-laden moral rebukes may allow human teammates to cultivate a virtue of reciprocity (*shu*, 恕) and the "heart of shame" (*xiuwu zhixin*, 羞恶之心) that are crucial for Confucian self-cultivation. First, 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 (including robots) to humiliate them (Liu 2017). Liu (2017) argues that this Confucian *negative* Golden Rule is more effective than the Christian *positive* Golden Rule "Do unto others as you would have them do unto you" as "what people do not desire has more common ground than what people do desire." In this sense, if the human teammate requests the robot to humiliate or harm others, blame-laden moral rebukes might help the human teammate cultivate the virtue of reciprocity: the human teammate does not wish for others including the robot to mistreat herself.

Second, blame-laden moral rebukes may help the human teammate cultivate the heart of shame. Confucian moral psychology emphasizes the embodied moral mind and combines both the body (*shen*, 身) and the heart (*xin*, 心) (Seok 2013). Embodied emotion serves as the foundation of virtuous dispositions. In this sense, a person's engaged moral experience does not only involve deliberative moral reasoning, as emotion also plays a crucial role in Confucian moral psychology. Our moral approvals and disapprovals are often exhibited through the employment of the body. For instance, when the robot blames a human teammate for her inappropriate moral request, the human teammate's innate heart of shame may bring some embodied emotional reactions (e.g., red face, sweating, accelerated heart rate). These different levels or forms of embodied emotional reactions are crucial for the cultivation of the "heart of shame" which may be possible in the interactions between the robot and human teammate, especially when they have developed long-term, affective relationships. The heart of shame is considered by Mencius as one of the four innate ethical tendencies which can grow into the virtue of righteousness (*yi*, 义).

Neo-Confucians such as Wang Fuzhi advocated for the "timeliness" and "expected utility" of moral remonstrations and blame. First, the timeliness of when to blame is crucial for the quality of blame. When a slightly selfish desire arises, Wang suggested that that desire will recede after immediate blame (Huang 2007). 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. Second, the "expected utility" of blame is also important for determining whether blame is necessary or not. Confucians recommend that a friend blames a person only if it is expected that such blame helps a person be reflective about her own conduct and grow her own virtues and moral sensitivity.

This capability requires two key capabilities: first-order theory of mind, and empathy. These two topics have been the topic of significant research over the past two decades. First-order theory of mind, i.e., the modeling of other humans' mental states, has shown great promise within the field of human–robot interaction (Scassellati 2002), with approaches developed to enable robots to model teammates' mental states (Devin and Alami 2016), understand differences in knowledge between themselves and their teammates (Hiatt et al. 2011), and

understanding what teammates are attending to Nagai et al. (2003). Moreover, in our own research we have developed language understanding capabilities that leverage second-order theory of mind, i.e., modeling what others likely believe the robot to believe—or in our case, what others likely believe the robot to be focusing on, attending to, and so forth (Williams et al. 2016; Williams and Scheutz 2019). Empathy, a process typically cast in terms of successfully identifying and responding to affective cues (Feshbach 1987), has also received significant attention, with numerous models presented for simulating or emulating empathy (Leite et al. 2013; Pereira et al. 2010; Tapus and Mataric 2007).

Given these capabilities, it is possible to ascribe moral blame. In general, moral blame is more effective when the robot discovers the immoral desire or tendency in the request of the human teammate than when such immoral desire or tendency has already caused unwelcoming consequences. As the Master remarked, "One does not try to explain what is over and done with, one does not try to criticize what is already gone, and one does not try to censure that which is already past" (*Analects* 3:21). In this sense, from the perspective of moral cultivation, it is less effective to design a robot to blame the human teammate *after* the robot sees the human teammate commit a wrongdoing than provide timely blame-laden moral rebukes when the robot identifies the unethical intention in the human request *before* the wrongdoing is conducted.

Confucians distinguish the exemplary person (*junzi*, 君子) and the petty person (*xiaoren*, 小人) through their reactions to blame. The petty person seeks blame in others (Brindley 2009). Unlike the petty person, the exemplary person turns blame into an opportunity for self-cultivation. Therefore, a long-term or life-long project for the Confucian person is to shift the vehicle for moral development from robot-generated blame (via blame-laden moral rebukes) to opportunities for "self-blame" (wherein humans consciously interrogate their own behaviors). In this sense, with frequent and everyday interaction with the morally competent robot, which is capable of making blame-laden moral rebukes, the human teammate has the potential to cultivate the "heart of shame." Thus, such cultivation of the heart of shame has the potential to transform a person's shameful feeling to self-blaming (Seok 2013).

However, from Western perspectives, one potential challenge for Confucian robot ethics might be: how should a robot interact with others with whom they have not developed close relationships? Can these robots respond strangers' moral violations with blame-laden rebukes as they do to their human teammates? A crucial concept "care with distinction" can provide a possible response to such challenge. As pointed out by Bell and Metz (2011, p. 88):

> Our ethical obligations, at least with regard to beneficence, are strongest to those with whom we have personal relationships, and they diminish in intensity the farther we go from those relationships. We do have an obligation to extend love beyond intimates, but there is not the expectation that the same degree of emotions and responsibilities will extend to strangers. The web of caring obligations that binds family members is more demanding than that binding citizens (or perhaps legal residents), the web of such obligations that

> bind citizens is more demanding than that binding foreigners, the web binding humans is more demanding than that binding nonhuman forms of life, and so on (Bell and Metz 2011, p. 88).

Arguably, as discussed earlier, social robots are often designed for fulfilling certain purposes or roles. These purposes or roles are often realized in the interaction between the robots and their human teammates rather than strangers. In long term interaction with their human teammates, or what computer scientists would call "deep learning," social robots might be able to develop "moral knowledge" that can be transferable to other similar contexts (Confucians such as Mencius would call such moral knowledge transfer "the extension of love"). Nevertheless, the level of moral concerns about strangers would be lower than the level of concerns about their human teammates. Also, more familiar contexts (e.g., the relationships with human teammates) will be likely to provide more information and resources for social robots to make good judgments. While in this paper, we have specifically examined the utility of blame-laden moral rebukes, we again stress that a robot's choice of norm violation response strategy will ultimately need to vary based on environmental, social, and dialogue context. In fact, our ultimate goal is to determine what communication strategies—beyond blame-laden moral rebukes—may most effectively shift the vehicle for moral development.

Implications for Designing Morally Competent Robots

Finally, there are potential implications of the Confucian theories including the theories on moral disapprovals (e.g., remonstrations, blames, and rebukes) for designing socially integrated and morally competent robots. As robots are increasingly becoming teammates, friends, and companions, it is critical to reflect on what constitutes morally reliable human–robot interaction that can bring positive moral experience and moral development opportunities to human teammates. As we have discussed, to design morally competent robots is to create not only reliable and efficient human–robot interaction, but also a robot-mediated environment in which human teammates can grow their own virtues.

First, we argue that Confucian role-based ethics can help roboticists make "visible" the *relational character* of the "autonomy" of robots perceived by humans. Dumouchel and Damiano (2017) recently argue that social robots such as Geminoid and Paro can only truly interact with other agents, and not with objects. Unlike humans, these robots have no relation to the world but their human partners. These robots were mainly created for the interaction or relationship with human partners. It is the interaction or relationship between robots and their human partners that makes the *existence* of these robots. In this sense, we suggest that roboticists should not only leverage the traditional, dominant approaches to developing artificial moral agents (AMAs) that focus on integrating rule-based morality, but also consider an alternative approach to designing morally competent robots based on the *role responsibilities* prescribed by the relationships robots have with human teammates in specific use contexts. We argue that a robot's selection of specific strategies

(including the severities of these strategies) for responding to human violations of norms (e.g., polite or humorous responses or blame-laden rebukes) depends on the roles this robot assumes and the relationships this robot has with its human teammates in specific temporal and spatial contexts.

Second, roboticists need to provide better reasoning and justification for integrating standing norms (e.g., being polite) (Isaac and Bridewell 2017) into robots to make them *socially integrated* robots. So-called "standing norms" are norms or maxims that are expected to be followed during ordinary conversations. Typical examples of these standing norms including being kind, polite, clear, and honest. A crucial role of standing norms such as kindness is often *instrumental* for providing a human-friendly, effective environment that introduces "ulterior motives" or the special goals interpersonal conversations attempt to achieve (Isaac and Bridewell 2017). These standing norms are also crucial for ensuring that robots are socially integrated as they are often expected to be followed by *actual* humans in everyday interactions. When designing the social integratedness of robots, roboticists need to ask themselves whether and how standing norms such as being polite can better support rather than supersede ulterior motives, whether the politeness of robots has its own intrinsic value, and whether making robots too polite will render them untrustworthy, with what Confucius would call "a clever tongue and fine appearance."

Third, findings from Confucian role-based ethics may inspire research on the moral development of human teammates in their *everyday* interaction with linguistically capable robots. For example, as indicated in the last section, Confucian ethics assigns robots a role responsibility of caring about the moral development of their human teammates by employing blame-laden moral rebukes as a strategy for *activating* self-reflection in human teammates. One hypothesis generated from such an assumption is that robots using role-based moral communication strategies when generating responses to human requests (especially morally inappropriate ones) will increase moral reflection and mindfulness in their human teammates. Furthermore, compared to moral language grounded in rule-based, deontological moral theories, role-based language may make more indirect reference to violated norms (e.g., deontological moral theories may draw more attention to the seriousness of violating universally applicable principles), requiring listeners to exert additional cognitive processing to identify the norm violation to which the speaker is truly responding. This intentional reflection is a quintessential prerequisite of state mindfulness (Shapiro et al. 2006) and corresponds to moral self-reflection in Confucianism. From the Confucian perspective, assuming humans do not wish for others to humiliate them, further empirical studies are needed to examine whether blame-laden moral rebukes inspired by the Confucian negative Golden Rule will help develop self-reflection, the virtue of reciprocity, and the heart of shame in human teammates. Finally, it may be the case that blame-laden moral rebukes, even when couched in indirect role-based moral language, will lead to embodied moral reactions (e.g., red face, sweating, accelerated heart rate) in human teammates that cannot easily be captured by deliberative moral reasoning, yet which are crucial for effective cultivation of the Confucian moral self.

Nevertheless, as pointed out by two reviewers of this paper, the contexts of human relations or roles might be crucial for the effectiveness of blame-laden rebukes in activating self-reflection. First, as mentioned earlier, to a self-conscious

and reflective person, a polite indication might be more persuasive than a strong moral rebuke. From the Confucian perspective, given the power distance between the father and the son, moral rebukes may not be effective or encouraged in such relationship for the son, despite that there are very few realistic situations in which a robot plays the son's role. Therefore, the question then becomes: can social robots have the computational capability to recognize the differences between these different contexts that are crucial for the effectiveness of blame-laden rebukes?

Second, for the Confucian interpretation of robot ethics to be successful, robots need to be recognized as agents by their human teammates. In fact, this is exactly what has been repeatedly observed in the human–robot interaction literature. Specifically, research has shown that humans not only perceive robots as autonomous agents (multiple studies to this effect are detailed by Scheutz 2012), but perceive them as social agents (Nass et al. 1994; Simmons et al. 2011; Straub 2016) and as moral agents (Kahn et al. 2012; Malle et al. 2015), with the caveat that humans apply different moral norms and different moral frameworks to robotic agents than they do to human agents (Malle et al. 2015). In our own work (Jackson and Williams 2019b), we have highlighted how natural language capable robots fit a unique niche with respect to perceived agency, where natural language capabilities may lead humans to intuitively ascribe social moral agency, a status that comes with unique persuasive powers (Jackson et al. 2019; Briggs and Scheutz 2014; Kennedy et al. 2014).

In conclusion, we hope this paper can help start the conversation among researchers in both robotics and applied ethics on how to understand the influence of robot communication strategies on the moral development of human teammates. As we have emphasized multiple times in this paper, a truly socially integrated robot, from the perspective of Confucian personhood, can and should have an ability to contribute to the development of a flourishing moral ecology in a society mediated by robots.

Acknowledgements This work was funded in part by National Science Foundation grant IIS-1909847.

References

Ames, R. T. (2011). *Confucian role ethics: A vocabulary*. Hong Kong: The Chinese University of Hong Kong Press.

Ames, R. T. (2016). Theorizing "person" in Confucian ethics: A good place to start. *Sungkyun Journal of East Asian Studies, 16*(2), 141–162.

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

Bell, D., & Metz, T. (2011). Confucianism and Ubuntu: Reflections on a dialogue between Chinese and African traditions. *Journal of Chinese Philosophy, 38*(s1), 78–95.

Brandstetter, J., Beckner, C., Sandoval, E. B., & Bartneck, C. (2017, March). *Persistent lexical entrainment in HRI*. Paper presented at the 2017 ACM/IEEE international conference on human–robot interaction, Vienna, Austria. Retrieved from https://dl.acm.org/citation.cfm?id=3020257

Brindley, E. (2009). "Why use an ox-cleaver to carve a chicken?" The sociology of the ideal in the Lunyu. *Philosophy East & West, 59*(1), 47–70.

Briggs, G., & Scheutz, M. (2014). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics, 6*(3), 343–355.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge, UK: University Press.

Burgoon, M., Hunsaker, F. G., & Dawson, E. J. (1994). *Human communication* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Confucius. (2000). *Confucius Analects: With selection from traditional commentaries* (trans by Slingerland, E.). Indianapolis, IN: Hackett.

Cormier, D., Newman, G., Nakane, M., Young, J. E., & Durocher, S. (2013, August). *Would you do as a robot commands? An obedience study for human–robot interaction*. Paper presented at the First international conference on human–agent interaction, Sapporo, Japan. Retrieved from https://hci.cs.umanitoba.ca/assets/publication_files/2013-would-you-do-as-a-robot-commands.pdf

Devin, S., & Alami, R. (2016, March). *An implemented theory of mind to improve human–robot shared plans execution*. Paper presented at the 11th ACM/IEEE international conference on human robot interaction, Christchurch, New Zealand. Retrieved from https://ieeexplore.ieee.org/abstract/document/7451768/

Dumouchel, P., & Damiano, L. (2017). *Living with robots* (trans, DeBevoise, M.) Cambridge, MA: Harvard University Press.

Feshbach, N. D. (1987). Parental empathy and child adjustment/maladjustment. In N. Eisenberg & J. Strayer (Eds.), *Cambridge studies in social and emotional development. Empathy and its development* (pp. 271–291). New York, NY: Cambridge University Press.

Goette, L., Huffman, D., & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review, 96*(2), 212–216.

Häring, M., Kuchenbrandt, D., & André, E. (2014, March). *Would you like to play with me? How robots' group membership and task features influence human–robot interaction*. Paper presented at the 2014 ACM/IEEE international conference on human–robot interaction, Bielefeld, Germany. Retrieved from https://dl.acm.org/citation.cfm?id=2559673

Hiatt, L. M., Harrison, A. M., & Trafton, G. J. (2011, July). *Accommodating human variability in human–robot teams through theory of mind*. Paper presented at the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain. Retrieved from https://dl.acm.org/citation.cfm?id=2283745

Huang, Y. (2007). Is Wang Yangming's notion of innate moral knowledge (liangzhi) tenable? In V. Shen & K.-L. Shun (Eds.), *Confucian ethics in retrospect and prospect* (pp. 149–170). Washington, DC: The Council for Research in Values and Philosophy.

Iio, T., Shiomi, M., Shinozawa, K., Miyashita, T., Akimoto, T., & Hagita, N. (2009, October). *Lexical entrainment in human–robot interaction: can robots entrain human vocabulary?* Paper presented at the 2009 IEEE/RSJ international conference on intelligent robots and systems, St. Louis, MO. Retrieved from https://ieeexplore.ieee.org/abstract/document/5354149

Isaac, A. M., & Bridewell, W. (2017). White lies on silver tongues: Why robots need to receive (and how). In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 155–172). New York, NY: Oxford University Press.

Jackson, R. B., & Williams, T. (2018, August). *Robot: Asker of questions and changer of norms?* Paper presented at the international conference on robot ethics and standards, Troy, NY. Retrieved from https://inside.mines.edu/~twilliams/pdfs/jackson2018icres.pdf

Jackson, R. B., & Williams, T. (2019a, March). *Language-capable robots may inadvertently weaken human moral norms*. Paper presented at the 14th ACM/IEEE international conference on human–robot interaction, Daegu, South Korea. Retrieved from https://inside.mines.edu/~twilliams/pdfs/jackson2019althri.pdf

Jackson, R. B., & Williams, T. (2019b, March). *On perceived social and moral agency in natural language capable robots*. Paper presented at the HRI workshop on the dark side of human–robot interaction: Ethical considerations and community guidelines for the field of HRI, Daegu, South Korea.

Jackson, R. B., Wen, R., & Williams, T. (2019, January). *Tact in noncompliance: The need for pragmatically apt responses to unethical commands*. Paper presented at the AAAI/ACM conference

on artificial intelligence, ethics, and society, Honolulu, HI. Retrieved from https://inside.mines.edu/~twilliams/pdfs/jackson2019aies.pdf

Jehn, K. A. (1997). A qualitative analysis of conflict types and dimensions in organizational groups. *Administrative Science Quarterly, 42*(3), 530–557.

Jung, M. F., Martelaro, N., & Hinds, P. J. (2015, March). *Using robots to moderate team conflict: The case of repairing violations*. Paper presented at the 10th Annual ACM/IEEE international conference on human–robot interaction, Portland, OR. Retrieved from https://dl.acm.org/citation.cfm?id=2696460

Kadar, D., & Marquez-Reiter, R. (2015). (Im)politeness and (im)morality: Insights from intervention. *Journal of Politeness Research Language Behaviour Culture, 11*(2), 239–260.

Kahn, P., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., et al. (2012, March). *Do people hold a humanoid robot morally accountable for the harm it causes?* Paper presented at the 7th ACM/IEEE international conference on human–robot interaction, Boston, MA. Retrieved from https://ieeexplore.ieee.org/abstract/document/6249577/

Kennedy, J., Baxter, P., & Belpaeme, T. (2014, March). *Children comply with a robot's indirect requests*. Paper presented at the 2014 ACM/IEEE international conference on human–robot interaction, Bielefeld, Germany. Retrieved from https://dl.acm.org/citation.cfm?id=2559636.2559820

Korsgaard, C. M. (1993). The reasons we can share: An attack on the distinction between agent-relative and agent neutral values. *Social Philosophy and Policy, 10*(1), 24–51.

Kronrod, A., Grinstein, A., & Wathieu, L. (2012). Go green! Should environmental messages be so assertive? *Journal of Marketing, 76*(1), 95–102.

Lakoff, R. T., & Ide, S. (Eds.). (2005). *Broadening the horizon of linguistic politeness*. Amsterdam, Netherlands: John Benjamins Publishing.

Lambert, A. (2017). Impartiality, close friendship and the Confucian tradition. In C. Risseeuw & M. van Raalte (Eds.), *Conceptualizing friendship in time and place* (pp. 205–228). Amsterdam, Netherlands: Brill.

Lee, N., Kim, J., Kim, E., & Kwon, O. (2017). The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *International Journal of Social Robotics, 9*(5), 727–743.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human–robot relations. *International Journal of Human-Computer Studies, 71*(3), 250–260.

Liu, J. (2017, October). *Confucian robotic ethics*. Paper presented at the international conference on the relevance of the classics under the conditions of modernity: humanity and science. Hong Kong: The Hong Kong Polytechnic University.

Lopez, A., Ccasane, B., Paredes, R., & Cuellar, F. (2017, March). *Effects of using indirect language by a robot to change human attitudes*. Paper presented at the 2017 ACM/IEEE international conference on human–robot interaction, Vienne, Austria. Retrieved from https://dl.acm.org/citation.cfm?id=3029798.3038310

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). *Sacrifice one for the good of many? People apply different moral norms to human and robot agents*. Paper presented at the ACM/IEEE international conference on human–robot interaction, Portland, OR. Retrieved from https://dl.acm.org/citation.cfm?id=2696458

Midden, C., & Ham, J. (2012). The illusion of agency: The influence of the agency of an artificial agent on its persuasive power. In M. Bang & E. L. Ragnemalm (Eds.), *Persuasive technology design for health and safety: Proceedings of the 7th international conference on persuasive technology* (pp. 90–99). Heidelberg, Germany: Springer.

Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., & Hagita, N. (2009, March). *Footing in human–robot conversations: How robots might shape participant roles using gaze cues*. Paper presented at the 4th ACM/IEEE international conference on human–robot interaction (HRI), La Jolla, CA. Retrieved from https://ieeexplore.ieee.org/document/6256095/

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science, 15*(4), 211–229.

Nass, C., Steuer, J., & Tauber, E. R. (1994, April). *Computers are social actors*. Paper presented at the SIGCHI conference on human factors in computing systems, Boston, MA. Retrieved from https://dl.acm.org/citation.cfm?doid=191666.191703

Nuyen, A. T. (2007). Confucian ethics as role-based ethics. *International Philosophical Quarterly, 47*(3), 315–328.

Pereira, A., Leite, A., Mascarenhas, S., Martinho, C., & Paiva, A. (2010, May). *Using empathy to improve human–robot relationships*. Paper presented at the 9th international conference on autonomous agents and multiagent systems, Toronto, Canada. Retrieved from https://dl.acm.org/citation.cfm?id=1838194

Puett, M., & Gross-Loh, C. (2016). *The path: What Chinese philosophers can teach us about the good life*. New York, NY: Simon & Schuster Inc.

Randall, T. E. (2019). Justifying partiality in care ethics. *Res Public*. https://doi.org/10.1007/s11158-019-09416-5.

Rea, D. J., Geiskkovitch, D., & Young, J. E. (2017, March). *Wizard of awwws: Exploring psychological impact on the researchers in social HRI experiments*. Paper presented at the 2017 ACM/IEEE international conference on human–robot interaction, Vienna, Austria. Retrieved from https://dl.acm.org/citation.cfm?id=3034782

Rosemont, H., & Ames, R. T. (2016). *Confucian role ethics: A moral vision for the 21st century?*. Taipei, Taiwan: National Taiwan University Press.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots, 12*(1), 13–24.

Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, K. Abney & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 205–222). Cambridge, MA: The MIT Press.

Seddon, K. H. (n.d.). Epictetus. *International encyclopedia of philosophy*. https://www.iep.utm.edu/epictetu/. Accessed 12 April 2019.

Seok, B. (2013). *Embodied moral psychology and Confucian philosophy*. Lanham, MD: Lexington Books.

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., et al. (2017). The case of classroom robots: teachers' deliberations on the ethical tensions. *AI & Society, 32*(4), 613–631.

Shapiro, S. L., Carlson, L. E., Astin, J. A., & Freedman, B. (2006). Mechanisms of mindfulness. *Journal of Clinical Psychology*, *62*(3), 373–386.

Shen, S., Slovak, P., & Jung, M. F. (2018, March). *"Stop. I see a conflict happening.": A robot mediator for young children's interpersonal conflict resolution*. Paper presented at the 2018 ACM/IEEE international conference on human–robot interaction, Chicago, IL. Retrieved from https://dl.acm.org/citation.cfm?id=3171248

Siegel, M., Breazeal, C., & Norton, M. I. (2009, October). *Persuasive robotics: The influence of robot gender on human behavior*. Paper presented at 2009 IEEE/RSJ international conference on intelligent robots and systems, St. Louis, MO. Retrieved from https://ieeexplore.ieee.org/document/5354116

Simmons, R., Makatchev, M., Kirby, R., Lee, M. K., et al. (2011). Believable robot characters. *AI Magazine, 32*(4), 39–52.

Straub, I. (2016). "It looks like a human!" the interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription. *AI & Society, 31*(4), 553–571.

Tapus, A., & Mataric, M. (2007, March). *Emulating empathy in socially assistive robotics*. Paper presented at the AAAI spring symposium on multidisciplinary collaboration for socially assistive robotics, Palo Alto, CA. Retrieved from https://robotics.usc.edu/publications/media/uploads/pubs/533.pdf

Tsuzuki, M., Miyamoto, S., & Zhang, Q. (1999). *Politeness degree of imperative and question request expressions: Japanese, English, Chinese*. Paper presented at the 6th international colloquium on cognitive science, Tokyo, Japan.

Vollmer, A.-L., Rohlfing, K. J., Wrede, B., & Cangelosi, A. (2015). Alignment to the actions of a robot. *International Journal of Social Robotics, 7*(2), 241–252.

Vollmer, A.-L., Wrede, B., Rohlfing, K. J., & Cangelosi, A. (2013, August). *Do beliefs about a robot's capabilities influence alignment to its actions?* Paper presented at the IEEE 3rd joint international conference on development and learning and epigenetic robotics (ICDL), Osaka, Japan. Retrieved from https://ieeexplore.ieee.org/document/6652521

Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016, March). *Situated open world reference resolution for human–robot dialogue*. Paper presented at the 11th ACM/IEEE international conference on human–robot interaction, Christchurch, New Zealand. Retrieved from https://ieeexplore.ieee.org/document/7451767

Williams, T., Jackson, R. B., & Lockshin, J. (2018, July). *A Bayesian analysis of moral norm malleability during clarification dialogues*. Paper presented at the 40th Annual Meeting of the Cognitive Science Society, Madison, WI. Retrieved from https://inside.mines.edu/~twilliams/pdfs/williams2018cogsci.pdf

Williams, T., & Scheutz, M. (2019). Reference in robotics: A givenness hierarchy theoretic approach. In J. Gundel & B. Abbott (Eds.), *Oxford handbook of reference*. Oxford, UK: Oxford University Press.

Wong, D. B. (2014). Cultivating the self in concert with others. In A. Olberding (Ed.), *Dao companion to the Analects* (pp. 171–198). Dordrecht, Netherlands: Springer.
Wong, P.-H. (2012). Dao, harmony and personhood: Towards a Confucian ethics of technology. *Philosophy and Technology, 25*(1), 67–86.