Robots, Rebukes, and Relationships:
Confucian Ethics and the Study of Human-Robot Interactions

Alexis Elder

(This is a pre-proof draft manuscript. Please cite the published version, in *Res Philosophica*, at https://doi.org/10.11612/resphil.2261)

Introduction

Studies of human-robot interaction (HRI) offer us a unique opportunity to dig into the details of moral psychology. Anthropomorphic appearances combined with fine-grained control over how, when, and whether they provide social stimuli can let us get a clearer picture of how people engage in social settings than would be possible using human agents, given the messiness of person-to-person interactions. Doing so can answer questions that are important for effective robot design, like "what social cues facilitate or inhibit human-robot collaboration at task *t*?" or "what implicit norms affect user experience of interaction *i*?". At the same time, interpretation of empirical data occurs within the context of particular values and theoretical frameworks, and theory considerations can drive decisions about how we should interpret data.

In what follows, I use empirical research on human shame responses in human-robot interaction to investigate questions about the ways in which shame can be triggered, and what functions it can serve in moral development. I situate these findings in the context of Confucian discussions of shame as a foundation of ethical virtue, arguing that this makes sense of extant data, helps further these same discussions by clarifying the factors contributing to effective ethical shame response, and suggests promising directions for design of robot-human interactions as tools for moral development.

Shame is interesting both in moral psychology and for global work in philosophy. It is widely recognized to be unpleasant to experience. But there is less agreement about its value; in particular, its contributions to moral development are contested. In both European and American philosophical and psychological literatures, shame is presumed to be a generally harmful, destructive, or broadly ‘bad’ emotion, one we would be better off without, or at least with a good deal less of it (Barrett 2015, Seok 2017). But in Confucian philosophy, shame is considered to be a “sprout”, or emotional foundation, of one of the four central virtues: our capacity to feel shame is what allows us, given appropriate cultivation, to develop the mature virtue of righteousness. Although we may initially feel shame at some things that we ought not, like wearing cheap clothing, and fail to feel shame when we ought to, like when bending a rule for oneself, we can with time, practice, and guidance become adept at scrutinizing ourselves to determine whether our conduct appropriately warrants shame. Importantly, shame involves openness to learning from others and a counter to one’s own tendencies toward hubris as one refines an appropriate sense of what is truly shameful, presenting us with opportunities for growth in social contexts (van Norden 2002). Modern research on psychology in East Asian communities finds shame to be associated with moral development and social connection (Seok 2015), as will be discussed in greater detail in subsequent sections.

As we will see, understanding how and why different philosophical traditions reach the conclusions they do may depend on fine-grained details about the ways in which shame can be experienced. I begin with an account that involves empirical research that appears to support the Confucian approach, where shame plays a useful and positive role in people’s moral development, before raising some apparently contradictory results from adjacent work on how shame impacts learning in human-robot interactions. I identify a way in which the different seemingly conflicting results can be consistently integrated into a unified and consistent theory, one which supports shame's potential positive contributions but identifies important risks and qualifications.

Note that I will be using the term 'robot' in a broad sense, to include both physically embodied mechanical entities and artificial intelligence-powered 'bots' that perform at least some human-like roles. Giving clear and precise definitions of exactly what constitutes a robot is quite difficult, and nothing substantive about this project hangs on a given definition.

Blame and Shame in Human-Robot Interaction

To err is human, so we need strategies to respond to our errors. What exactly this should look like seems to depend in part on empirical questions about what kinds of responses best contribute to our moral development.

In "Blame‐Laden Moral Rebukes and the Morally Competent Robot: A Confucian Ethical Perspective", Qin Zhu, Tom Williams, Blake Jackson, and Ruchen Wen survey literature on social robots' capacity to influence human moral decision-making, and argue for an increased focus on developing robots' capacities to issue context-appropriate moral rebukes, in order to contribute to what they call a flourishing "moral ecosystem". Against presumptions that people would object to or ignore robotic rebukes, they discovered that when robots appear to share and express moral values with human collaborators, people both found these robots more likable and were better able to adhere to their own moral commitments.

They begin by describing an experiment in which human subjects were asked to consider performing an immoral action, and then shown video footage and/or dialogues in which robots responded to commands to perform these same actions by asking questions addressing the practical implementation of the command but not its permissibility, such as which of two computers to destroy (2531). After these interactions, subjects' responses showed both that they interpreted the robots' ignoring the ethical dimensions of the task as tacit evidence that the robot viewed the action as permissible, and furthermore, shifted their own stance on its immorality, viewing it as more acceptable than they did before. That is, rather than read the technology as morally neutral, people were likely to read the neutral response as a value judgment undermining the apparent impermissibility of the task.

They argue that this shows we do not, by default, consider anthropomorphic robots to be outside our moral ecosystem, but need to consider their place within it. "[I]f 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", they caution. (2513) Furthermore, if robots *do* react negatively to norm violations, this can increase human trust and acceptance of them. "In our own work," they note, "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." (2514) That is, not only is it *morally* valuable to have robots capable of strengthening human moral commitments, it is *pragmatically* desirable from a design perspective to build robots that human collaborators regard as more likable and appropriate, and this in turn can amplify their persuasive power when it comes to moral influence, creating a virtuous feedback loop.

The importance of character versus situations in influencing moral behavior is a familiar one to moral psychologists. Among moral psychologists, one relatively long-standing debate concerns the extent to which we should explain moral behavior in terms of character traits of individuals versus situational influences. For example, is a particular altruistic gesture better understood as an expression of the agent’s generous disposition, or a response to environmental cues prompting people to contribute? The team’s findings would be interesting even if one were interested in a strictly situationist assessment of robots’ impacts on human actions here, valuing robots merely for their capacity to improve human moral behavior when sharing environments with them. Just as situationists in discussions of moral psychology will point to evidence about the salutary effects of finding dimes in phone booths on increasing altruistic behavior (Doris 1998), one might take this to be an interesting effect to be integrated into a situationist framework, useful in constructing contexts for human-robot collaboration that would be more likely to help human beings act ethically. But Zhu et al use the opportunity to investigate robots’ potential to cultivate character traits in a more virtue-theoretic framework, albeit one that differs significantly from the Aristotelian kind more commonly found in situationist/character ethics debates. They turn to an account of long term human moral development via the kind of norm internalization found in Confucian accounts of the relationship between rebuking practices in relationships and the development of people’s “hearts of shame.”

They begin by emphasizing the relational account of personhood that they find to be characteristic of Confucianism, highlighting its emphasis on role ethics, such that robots ought not be thought of as merely presenting as person-*like* in general but inhabiting specific roles (teacher, friend, etc.) in relation to their human partners.

This may be a somewhat controversial assumption, as one important issue in Confucian ethics involves *zhengming*, “rectification of names”. The ‘names’ here involve social roles, like *ruler* and *child*, and ‘rectification’ involves both acting in ways that are appropriate to one’s role (the child should display filial piety) but also appropriate application of role terms (one should call one’s father ‘father’ rather than ‘brother’). If robots are not people, one might think, it would not be good to cast them in personal roles. However, Confucian conceptions of personhood are themselves relational and developmental, and it has been argued that one cannot fully understand Confucian personhood independent of roles (Nuyen 2009), so to rule them out as role candidates *a priori* might be to beg the question. Given that people already seem to take robots to be person-like participants in our ‘moral ecology’, then, I think it is worth exploring what happens when we investigate their occupation of social roles. Even if we end up concluding that robots are not the right kind of thing to fully occupy social roles for us, it can be useful to see what happens when they inhabit cartoonish or iconic forms of social roles.

These roles can come with distinctive expectations around norm adherence and response to norm violations. Teachers, for example, may be expected to pay special attention to students’ adherence to norms relevant to their field of instruction, whether linguistic (language classes), logical (in philosophy classes) or stylistic (studio arts), and to respond to violations of norms by correcting students for norm violations, as part of the project of helping students to internalize and refine their capacity to act in accordance with these norms. While it can be easier to see the significance of responding to norm violations in some roles, like that of teacher or parent, than others that are less formal or hierarchical, like friendship, Zhu et al plausibly make the case that even friendship calls for friends to be willing to rebuke each other for moral norm violations, despite these rebukes’ potential unpleasantness and even, in some cases, when doing so violates politeness norms.

This Confucian framework accords with people’s tendency to interpret even neutral robot responses within a moral framework, they argue, and directs us to consider the *kind* of role people experience robots as occupying in their interactions:

> ...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. (2516)

For example, when robots are viewed as fulfilling friend roles, the team predicts, their rebukes may be interpreted within the context of friendship, stimulating both immediate bodily responses to emotional cues designed to elicit human emotional response, and contributing to internalization of the critique via the experienced bodily reaction in conjunction with the friendly, trusting, affectionate and supportive framework in which it is presented.

> 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. (2517)

In particular, when robots are designed to provide timely and even preemptive rebukes that occur during or just before the person violates a moral norm (2519), that can activate the humans' own moral sensibilities. Especially in the context of longstanding relationships in which affection and felt trust are cultivated, these rebukes may have the effect of not just behaviorally influencing the person's actions when under the robot's supervision, but shaping their moral development and capacity to feel shame at their own actions, leading to greater long term propensity to act morally, consistent with Confucian accounts of the development of experienced shame into the virtuous disposition to have an appropriate sense of shame when contemplating wrongdoing:

> "...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

> team-mate 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).” (2519)

While speculative, this is a highly interesting suggestion, and one that seems amenable to empirical testing: can robots help people to cultivate an internalized “heart of shame” by reinforcing moral norms during collaboration?

The Structure of a Confucian Approach to Shame-Cultivating HRI

Given the importance of roles in identifying and articulating responsibilities, as well as the importance of trust and affection in creating conditions appropriate for cultivating and internal capacity for self-blame, they argue that this work is best done with robots occupying friend-like sustained relationships with the human beings they rebuke.

People show predispositions to respond to robots as moral agents such that even non-engagement with ethical issues is interpreted as having ethical valence (like implicit permission to engage in otherwise impermissible behavior), and rebukes and praise influence human decision making in ways that resemble interpersonal influences on moral behavior. It also seems plausible that people can have more or less friendly, long-term, trusting, affectionate relationships with robots in ways that resemble interpersonal relationships like friendship. For example, soldiers have formed bonds with their battlefield robots, expressing strong emotional attachments to them and even holding funerals for them when they are destroyed (Garber 2013).

A Confucian framework connects these two points by way of a relational account of personhood as inherently connected to occupation of roles we fulfill for each other. Confucianism also highlights the importance of moral ecosystems in shaping our moral capacities both short- and long term, in the short term by susceptibility to the approval or blame of others (and tacit influence of their own decision-making), and in the long term by internalizing blame, developing the capacity to feel shame at wrongdoing even in the absence of others, subjecting oneself and one's reasoning to scrutiny, perhaps – although, as we will see, accounts differ on this – as if observed and judged by others.

Furthermore, within a Confucian framework, rebukes and training to feel shame occur in the context of relationships, especially ones where trust and affection are involved, creating the conditions for people to feel the bodily and emotional experience of shame at wrongdoing rather than mere occasion for temporary embarrassment upon discovering that one has committed a social faux pas, thus coming to associate wrongdoing with this felt shame and leading to increased tendency to self-scrutinize and reflect on one's own behavior.

So it would seem that insofar as people respond to robots' disapproval by reflecting on their own behavior, Confucianism both aligns with empirical data on how people actually react to robots' judgment and suggests a strategy for putting it to work to promote moral flourishing (although the long term effects would need to be empirically confirmed).

This also gives us a case study in the ways that moral philosophy and technological research and design can be fruitfully intertwined, offering both explanations for extant data and suggesting future directions to explore.

Challenges from Educational HRI

But here is a potential problem for the account. The above focuses on the ways that robots' being experientially *person-like* (and, furthermore, occupying personish roles for human beings) can be morally significant. However, there is also evidence from studies of human-robot interaction that suggests that experiential *differences* between interacting with robots and people can have implications for human learning and development, and not just in general but particularly with respect to the work that felt shame in the face of the judgment of others does in Confucian theorizing. This presents a potential set of counter-examples not just to Zhu et al's project, but to the Confucian account of the developmental role of feeling shame before others in becoming a mature moral agent.

We have reason to think that robots can provide human-like assistance while at the same time providing a distinctive benefit by offering freedom from the perceived judgment of others, thus reducing activation of shame., In geriatric care technologies, for example, seniors may prefer robotic to human assistance with care tasks associated with shameful or embarrassing (but not immoral) activities, such as toileting and bathing, and see robots as valuable because they enable seniors to feel more dignified (Felber et al 2022). Where shame is a risk, robots' ability to present at a psychic distance seems significant in understanding how they compare to human agents.

One might worry about the relevance of the examples I am about to give, because the following cases involve learning but not learning *moral* topics. But they present an interesting challenge because they let us tease out the significance of occupation of a *role* in a relationship with people as distinct from appearing to people as other full-blown *people* whose judgment matters and thus (Confucian accounts predict) should help us to internalize an appropriate sense of shame. If these robots seem to occupy roles but not 'read' as persons, this seems to speak against the importance of interpersonally-stimulated felt shame in internalizing norms and learning to hold oneself accountable to them. Because of their value in exploring this distinction, and because shame generally can be felt at many things, moral and non-moral, from poor clothing to etiquette lapses to moral failures, I think they are worth investigating.

The first example involves the use of artificial intelligence to give feedback on paper drafts in educational contexts, as reported in a *Hechinger Report* story on research involving the effects of robot marking (Paul 2014). This is not, to be clear, about robot *grading*, but rather the use of robots to provide developmental feedback as students revise their paper drafts. In writing instruction, one important part of the process is giving students feedback on drafts. Students are then supposed to incorporate this feedback into revisions, rewriting and restructuring portions of their papers. To put this in terms friendly to the Confucian framework already introduced, students are expected to change their work in response to the judgment of experienced writing teachers and thus learn to internalize the standards of the more advanced writer in order to incorporate them into their own writing in future.

One ongoing challenge in this portion of the writing instruction process is students' tendency to *personalize* critical feedback, interpreting it as based in interpersonal dislikes and resentments in the context of the teacher-student relationship, and doing minimal work to improve their drafts (measured quantitatively as volume of rewriting performed), even though more extensive rewriting is associated with better long term performance at writing. Roughly, there is a straightforward interpretation whereby criticism feels like a personal attack, and students double down in the face of attack with resentment and defensiveness. When, however, they receive feedback from (what they perceive to be) impersonal automated robot readers, they engage in much more extensive revising (again, measured in terms of volume of new text produced) and greater enthusiasm for progressing as well as less resentment and resistance, treating it as a game to be mastered – how can this paper be revised so as to "win" full credit? (Paul 2014)

Thus, this example and related research seems to support the hypothesis that role occupation without personal attribution protects against a shame reaction that actually interferes with learning (Howley et al 2014).

The second potential counter-example involves language learning. One major challenge in second-language instruction is students' reluctance to make mistakes in front of others. This can make it difficult both for students to stretch themselves and take risks, for fear of interpersonal embarrassment, and actually make it difficult for formative assessment to take place, that is, for instructors to see where students are struggling, since they tend to avoid speaking in contexts where they fear they might err. Here as in the previous example, robots have proved helpful when they occupy quasi-instructor and assessment roles (Leyzberg et al 2018). Students paired with a (physically embodied) robot tutor showed a higher early incidence of mistakes as well as significant improvement in the longer term, and researchers hypothesized that these two factors may be causally related; willingness to make mistakes and to ask for help seem to be stronger in the presence of robots versus human instructors, and making mistakes is critical for learning (Leyzberg et al 2018).

Here again, shame's role in producing a fear of judgment and embarrassment in learners seemed to present an obstacle to the kind of practice necessary for learners to internalize standards, and robots, which occupy an in-between space between person and thing, seem to be able to fill person-like roles for students without presenting as persons capable of judgment. This seems to do important work in helping learners to internalize norms *without* activating the appearance of blame and judgment associated with Confucian accounts of such internalization, and furthermore to be more successful than human teachers *because* they present a kind of ‘judgment-free’ experience that nevertheless offers guidance and corrections necessary to development.

Thus, we get a unique challenge to Confucian accounts of the cultivation of something like the ‘heart of shame’ that is made possible by empirical work in human-robot interaction, where their status as person-like non-persons and their ability to occupy instructional roles lets us dig into the plausibility of the mechanisms by which people use rebukes and corrections to internalize norms, and in particular the distinctive bodily experience of shame before others’ judgment as opposed to the feeling of being in a ‘judgment-free’ zone that nevertheless includes the feedback and corrections normally associated with interpersonal interactions with human occupiers of social roles like those of teachers.

Some Possible Responses

This does not mean, however, that Confucian accounts of the cultivation of shame are thus refuted. I will explore two strategies of response available here before showing why I think that for a complete account, we need to dig deeper into the details of shame.

One strategy is to identify a difference between the role shame plays in non-moral activities and learning, and the role it plays in distinctively moral development. This would be to grant that shame before human judgment is in fact an obstacle to internalization of norms in non-moral cases, and thus positions robots as better candidates for teaching in these tasks while preserving space for distinctively human instruction in moral areas of learning. At the same time, it would involve postulating a different role for learning in moral contexts, one in which the judgment of others plays a more important role in norm internalization and the capacity to self-assess, and thus one which has not been measured by the above cases. But this seems ad hoc without further defense, and so I will be setting it aside here.

Another strategy could be to take the above as evidence that it is not sufficient for appropriate learning that any random *person* fulfill a role, but rather underscores the importance of trust and affection in interpersonal learning. This response might run something like: robots are better than typical human instructors at providing feedback in contemporary school settings with low levels of trust and affection between teachers and students, but this merely highlights a shortcoming of extant teacher-student relationships. That is, if writing students were not predisposed to interpret writing instructors' feedback as personal attacks, said predisposition itself being a reflection of antagonistic or flawed teacher-student relationships more generally, we might not see improvement in response to robot graders, because defensiveness in response to shame would not be triggered.(Note that this is distinct from claiming that the students actively *trust* or feel affection for the robots more than the teachers and thus have a better pre-existing relationship to provide context for rebukes. In fact, Zhu et al found that issuing rebukes was one of the things that helped *foster* affection toward robots. In any case, presumably the students in the study had stronger pre-existing relationships with their teachers than with the bots introduced for the study.)

One might think that students might be able to experience shame without defensiveness, or might not be so fearful of making mistakes when learning a new language if they felt *more* appropriate trust and affection for their human instructors while still valuing and internalizing their assessments and feedback, and perhaps these benefits would more accurately resemble the process described (in the ideal case) in Confucian discussions of the 'heart of shame'. I think the fact that in these experiments the robots can productively issue rebukes shows that more affectionate trust is not necessary for learning. But if we are to draw useful conclusions from these examples, we need to be sure we are comparing apples to apples, not merely identifying the harm done by flawed human relationships, said flaws already being accountable within a Confucian framework for understanding shame.

These apparent counter-examples muddy the waters, given Zhu et al's results. But I think there is a way that there is an explanation of how shame works that can unify and explain them all consistently. Doing so can shed light on both human-robot interactions and shame's role in moral development. But to do so we will need to go into a bit more detail about the structure and value of shame.

Deepening Our Understanding of Shame

To see what lessons can be drawn, given apparent conflict between these empirical investigations of human-robot interactions, it is helpful to look more closely at the details of several accounts of shame.

To start with, the status of shame within moral psychology is contested, along broadly cultural lines; it is often presumed in European and North American research that shame is destructive, inherently involves devaluation of the self, and activates denial and defensiveness (Barrett 2015, Seok 2015). Meanwhile, in many cultures across East Asia that are broadly associated with Confucian philosophical traditions, shame is considered to be positive, to promote moral self-reflection necessary for self-growth, and to be an important feature of close and caring relationships (Seok 2015). One possibility is to identify different cultural features that make different uses of shame in the context of varying cultural frameworks (Nichols 2015) That is, shame might be good for East Asians and bad for North Americans, because of differences in the cultures in which shame is activated and experienced. This might suggest that, for instance, robot rebukes would be appropriate for East Asian markets but not North American ones, or that instructional robots might have different effects on students in WEIRD (Western, Educated, Industrialized, Rich, and Democratic) vs non-WEIRD countries. (I am not aware of research on this but would be interested to see what results might look like.)

But I think we should not rush too quickly to this sort of cultural-relativist explanation. While it might seem tempting, especially if one associates this cultural difference with collectivist vs individualistic cultures, discussions about Confucian ethics and especially about shame itself give reason to proceed with care both in attributing a ‘pro-shame’ view to Confucianism and to conflating Confucian philosophy with East Asian culture generally, as well as relying too much on simplistic individualist vs. collectivist explanations. To start with, within the Confucian philosophical tradition (which, again, should be for clarity’s sake distinguished from the broader cultural contexts in which Confucianism has been incorporated), philosophers are very careful to distinguish between different forms of shame, and to recommend, not shame *simpliciter*, but appropriate cultivation of the *capacity* for shame at *appropriate* moral circumstances and not, for example, mere social mores or status signaling. In addition, philosophers have cautioned against overgeneralizing when it comes to collectivist vs. individualist explanations of philosophical differences *or* cultural practices (Olberding 2015, Wong 2008).

Lastly, as Seok (2015) and Van Norden (2002) argue, ‘shame’ itself is ambiguous between several different senses of the word, including the felt experience of shame, the nature of the object that elicits the experience (conventional vs. moral), the propensity to feel shame (in general, in response to particular objects that may be more or less justified in eliciting shame reactions), the behavior of others intended to produce felt shame, the use of shame as a motivator for self-reflection and change, etc., and Confucian texts offer quite detailed guidance on when, whether, why and how one *ought* to feel shame in various circumstances. For example, surveying Confucius' *Analects*, Van Norden notes that although people can feel *chi* (shame) about "poor clothes, poor food (Analect. 4.9), asking questions of social inferiors (5.15), and being poorly dressed in the presence of those who are well dressed (9.27)–they should not be *chi* about these things. We also learn that, whether they are or not, people should be *chi* about being poor and lowly in a well ordered state, or being wealthy and esteemed in an ill-ordered state (8.13), not living up to one's words (4:22), and toadying and feigning friendship (5.25)" (Van Norden 2002, 64).

Thus, Confucian philosophy does not present a defense of the value of feeling shame in general. We do not find endorsement of the idea that shame is something we ought to experience more frequently. Rather, accounts specifically defend the value of the *capacity* to feel shame *when one has done something that is morally wrong*, **and then** to be disposed to use that feeling as a motivator to reflect on where one falls short of one's own ideals and to work toward moving closer to these very ideals. The Confucian scholar Mengzi draws on a number of agricultural metaphors in his analysis of human moral psychology. In his framing, shame is the *sprout* of righteousness, not the full-grown plant, and just as not all feelings of alarm and commiseration are instances of mature benevolence and indeed they can be used in some circumstances to undermine it, neither is the feeling of shame (Van Norden 2002).

This aligns with the way that Zhu et al understand shame in their work on robot rebukes. Using this understanding, they go on to posit that shame works best in the context of close, trusting relationships, of the sort that can be occupied by either humans or robots. As we have seen, they identify affection, trust, and duration of relationships, as well as perceived likeability and appropriateness of the robot – which itself includes propensity to respond with appropriate severity to violations of moral norms – as better-making factors for people's internalizations of norms. They base their work on apparent similarities between new robotic occupants of established social roles, and more familiar human occupation of these same roles. And the results they discuss come from US participants, not East Asians (Jackson and Williams 2019). That is further reason to resist the temptation to reduce shame analyses to cultural influence.

As I said at the start, I am interested both in how philosophical theories can help us make sense of empirical evidence, and at the same time how empirical evidence can help us refine our philosophical theories. So far, I have focused on the first part, by drawing on Confucian resources to account for empirical evidence about shame responses in response to robot promptings. Contrary to the common Western trope that shame is bad and harmful to learning, Confucians identify a place for shame in both rich and rewarding interpersonal relationships and in motivating personal development by way of internalizing norms and spurring self-reflection and self improvement. But at the same time, evidence about the relative success of robots versus human beings at rebuking in teaching relationships teaches us something about how fragile the parameters for appropriate shame can be for developing people – the ability to work well with felt shame is itself a skill, and one that requires a great deal of trust. Given the flawed nature of many actual interpersonal relationships, one important route to developing the sprout of shame into mature norm internalization and capacity for self-reflection may involve activation of roles without the baggage of full blown persons inhabiting them. That is, the roles people inhabit do a great deal both to set expectations and provide guidelines for interaction that can help reduce interpersonal conflict. In addition, roles create contexts where people can learn to inhabit their roles to begin with. Shame is an especially good candidate to show this developmental work done by roles, as it is one where it is astonishingly easy to overshoot positive impact zones and activate a learner’s defensiveness and propensity to hide rather than grow. (This narrow zone for growth might be the sort of thing that can be widened with appropriate cultural practices and support, as Seok 2015 argues, but that is a topic outside the scope of the current project.)

Robots as Cartoons of People: Shame and Idealization

Zhu et al (2020) focus on making robots more person-like. But the results of other research into robot teaching shows that there is value in the cartoonish, iconic, stylized version of personhood found in robotics. Let me unpack this a bit. Readers may be familiar with the idea of the “uncanny valley,” in which human beings find robots likable when they resemble human beings… but not too much. (The “valley” refers to a dip in the otherwise-upward-trending line representing likability in a graph where the x axis represents degree of resemblance to actual humans, while the y axis represents, roughly, user preference and sense of familiarity.)

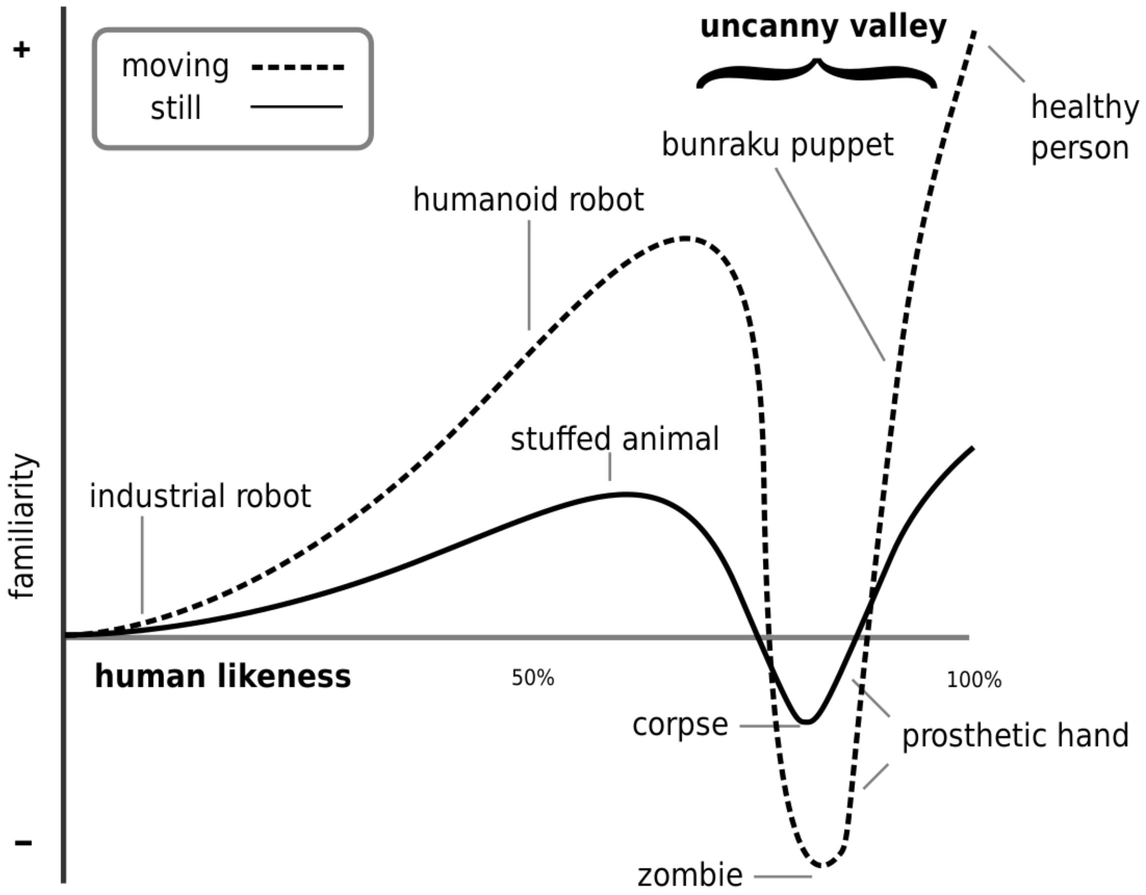

Figure 1: Uncanny Valley.
By Smurrayinchester – self-made, based on image by Masahiro Mori and Karl MacDorman at http://www.androidscience.com/theuncannyvalley/proceedings2005/uncannyvalley.html, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2041097

One explanation for this "valley" is that we find person-like things appealing, up to a point, but once an entity is person-like *enough* we interpret every difference as malicious (or at least worrying) – the emoji smiley-face is simple enough that we can just see it as a smile, while a fairly-realistic but not perfectly accurate AI-generated humanoid face can read as "off" or creepy because of small differences in feature placement or unorthodox movements of facial features that we interpret in an interpersonal framework as threatening or evidence of illness or dangerous deviance. Robot rebukes may function similarly. When we are provided with just the outlines of a role, we can accept them at face value and work with them as symbolic icons of the phenomenon, whereas with highly detailed interactions with complex humanoid figures (and perhaps even actual humans) we can find ourselves primed to read a great detail into even the smallest details, often to the detriment of our relationships, like students' interpretations of teacher feedback as hostile and interpersonally motivated rather than instructive. Cartoons may be useful because they do not provide enough detail to misinterpret.

This may be especially valuable with shame. In the Confucian framework, morally mature shame responses are not reactive to every actual human judgment of oneself. Rather, maturation of the shame response consists of developing the capacity to evaluate oneself from the perspective of internalized ideal observer. As Bongrae Seok puts it in a work comparing positive East Asian conceptions of shame with more negative Western European and American ones, the kind of shame we should aspire to has an interpersonal character but is heavily mediated by one's own standards (Seok 2015).

"From the viewpoint of the broad interpersonal, social, and moral dimensions of [morally positive] shame," writes Seok, "the reason one should be careful about one's own behavior is not because one's personal reputation is ruined by others' watching one's personal wrongdoings, but because one cares about one's whole self living a virtuous life in changing personal, social, and moral environments. Shame, in this sense, is a self-evaluative emotion, a constant process of reflective evaluation of oneself against one's moral ideal in the diverse and challenging conditions of human life." (37) But it is not *merely* measuring oneself against a standard and falling short – that would undersell the connection between mature moral shame and developmental shame before others. Even in the morally mature agent, argues Seok, shame occurs

> in the context of a moral agent's relation to anonymous others or to people (i.e., moral authority or norm) she respects. That is, when a person is ashamed, she feels an *inner sense* of violation *in front of* her moral ideal, manifested in the form of exemplary figures… This unique moral shame combines external shame (feeling ashamed in front of others) and internal shame (inner sense of morality) together in a unique and inclusive emotion of self-reflection and moral challenge. (38)

Even in its morally mature form, it still *feels* like shame before others, with its distinctive motivational impact, and its sense of something outside of oneself doing the assessing.

Cartoonish or iconic robots, with their pared-down but powerful appearance of personhood, may help us to achieve this experience as learners, in a way not available to messy, particular, fraught interpersonal relationships between human beings. This conception of mature shame highlights not just a developmental role for pared-down idealized others, in an instrumental attempt to blunt the effects of defensive reactions by the learner, but that the goal of shame cultivation itself is to relate to an idealized and anonymous imagined other, highlighting a unique opportunity for robots in particular to contribute to our moral development.

One way this might contribute to this development is by creating something like what Michael Puett calls an "as-if" space (Puett 2015), in which people can explore normative possibilities away from their immediate practical application, and that this playful exploration can be important for learning. (For a familiar example of this, consider how students are willing to do *more* writing and keep trying to improve scores when working with robotic essay-feedback systems.) Puett's account focuses on the ways that *li*, roughly, rituals and social norms like the rules of etiquette, can create breaks in thoughtless patterned response and invite us to step outside these well-worn grooves and try something different. Without endorsing his entire account of ritual or claiming that robots themselves constitute rituals, I am suggesting that a psychological function of engaging with rebuking robots might be to create a playful as-if space in which one can try on new norms (in the case of learning-assistance robots) or self-correction for considering violation of established norms (in the case of robots that issue moral rebukes at the suggestion of property damage) in a way that moves people outside of their ingrained and habituated patterns of interpersonal interaction.

Robots may create an as-if space for people to experience reflection and psychological distance, avoiding getting drawn unreflectively into damaging patterns of response. This would let people work more directly with ideals, at a remove from the messiness of high-stress interpersonal dynamics during critical development of skills, revealing the delicacy of shame capacity development work. This, in turn accords with Confucian emphasis on feeling shame at the right things rather than the wrong things that we have already seen, with the ultimate goal of shame experiences causing us to be reflective and attuned to value rather than reactive to social opinions, ultimately learning to internalize the capacity for self-reflection through the eyes of imagined exemplars rather than pursuing the approval of others.

The kind of playful as-if engagement here is distinct from modern conceptions of gamification. Gamification involves ascribing game-like rules and reward structures to a non-game, real-world application. To get clear on the difference, consider C. Thi Nguyen's account of the hazards of gamification. "Gamification increases our motivation by changing the nature of the activity", says Nguyen, in a criticism of the practice (Nguyen 2021, 411). "Often, the goals of ordinary activity are rich and subtle. When we gamify these activities, we change those goals to make them artificially clear." (411)

By contrast, I am suggesting that rather than overlay new values on an existing space, robots create a *new* space in which we can engage with *current* goals, values and norms, in order to explore and appreciate the opportunities afforded by their very richness and subtlety. This protects against an important objection to gamification that Nguyen raises: "In games proper, this simplification isn't particularly problematic, because …their associated goals, are usually kept secluded from ordinary life. But there is no such protective separation when we gamify ordinary activities" (411-12) The as-if space of robot-human interactions introduces such a protective separation, and furthermore allows us to grapple with the very real, potent, double-edged sword of shame as a motivator. By contrast, argues Nguyen, "[t]o reap the motivational benefits of gamification, we must reshape the ends which govern our real-life activities" (412). In gamification, we chase high scores and leaderboards. In robot-human rebukes, we delicately explore our own shame responses to violations of norms we already take to be valuable, to become more discerning and self-reflective when we feel ourselves motivated by shame as social creatures who benefit from responsiveness to others but need to modulate our reactivity.

Directions for Future Research

There remain many unanswered questions about shame, robots, and moral development. Clarifying philosophical theories of shame suggests directions for future empirical research that might be used to answer some of them, including the starting question: can robots help people to cultivate an internalized "heart of shame" by reinforcing moral norms during collaboration?

For example, HRI studies might investigate the Confucian idea that the social roles robots inhabit will affect how their rebukes are interpreted, both in the moment by measuring bodily responses like blushing that may be associated with the experience of shame, and in the long term as to whether and how the rebuke is internalized and carried forward by the human subject in future scenarios or post-experiment surveys. For example, researchers might try varying whether the robot is presented to the subjects well in advance so that they have time to become familiar with them, perhaps work together, and form affectionate bonds (which might also be measured at test time), in comparison to new robots introduced during the test as ‘strangers,’ or introducing robots as occupants of different social roles like ‘teacher’ or ‘colleague’, to see how this affects the way rebukes are received and what, if any, long-term differences result.

Both human and robot confederates might vary the extent to which they present as “non-judgmental” in learning contexts, to see whether and how species or judgmentalness affects shame reactions and long-term norm internalization by subjects. Comparing these results with students’ own tendencies to consider shame as a positive or negative emotion, and conducting this work across East Asian and North American cultural contexts, would also be helpful in understanding how judgment and shame are related to learning and the extent to which culturally-specific framings impact this relationship. In addition, to probe the hypothesized connection between cartoonish appearances and capacity for abstraction, researchers might try experimenting with details of robot appearances to see whether this impacts norm internalization, in particular whether people tend to ‘read’ more into human or near-human rebukes than robot ones. This could be accomplished by issuing the same rebuke with more or less realistic robots, or by comparing robot and human rebukes, and asking subjects for their interpretation of the rebuke as well as observing their performance post-rebuke.

The extent to which gamification might or might not be involved in contexts involving successful robot rebukes could be measured using measures of internal versus external motivation for activities following robot versus human confederate instruction or rebuke, as well as self-reported perceptions of shame and measures for self-reflection following the intervention or interventions. Finally, progress could be made in the area of applying findings to longer-term moral development by testing which features of human interaction show highest rates of transfer from HRI contexts to to human-to-human interactions, whether specifically focused on persistence of particular norm commitments, or receptivity to rebukes without damaging reactivity more generally.

Conclusion

At this point, I hope to have shown that, despite initial appearances, evidence from work with instructional robots do not show that shame is bad, in the Western sense. In fact, they show that correction or rebuke itself is not the problem, since learners are able to fruitfully engage with rebuking robots even where human rebukes interfere with learning. This is in accord with Zhu et al's finding that even vehement rebukes, when appropriate to norm significance and violation severity, actually *increase* robots' likeability, which shows that we actually value the capacity for issuing rebukes in our interpersonal relationships. Although a common narrative from Western psychology and ethics holds that guilt rather than shame is effective because guilt focuses on actions rather than selves, this is not supported by cross-cultural psychological research (Seok 2015). Instead, Confucian discussions of ideals of shame give us reason to think that shame is valuable when it helps us to develop our capacity for humble self-reflectiveness.

Research in human-robot interactions gives a unique opportunity to control for variables of human relationship and at the same time offer promise in guiding technological development with an eye toward creating what Zhu et al call "moral ecologies" where human agents can flourish. We can get fine-grained feedback on the mechanisms of the activation and development of capacities involving morally significant emotion. In this case, results seem to show that robots can help us to explore social roles and develop capacities while reducing the danger of felt shame driving us to hide or become unproductively defensive, derailing the interpersonal relationships that can ideally help us actually learn to work with and grow from experiences of shame.

In our interpersonal relationships, we imperfect human beings are always working with our own and others' imperfections, and one of the challenges of the human condition is figuring out how to occupy roles in relationships where various parties are imperfect at fulfilling their roles. This is a factor that Confucians are keenly aware of, as for example in discussions of the sage-king Shun's many struggles to practice filial piety with highly imperfect parents (Wong 2008). Robots let us explore roles in new ways, helping us to see the work done by roles versus persons while offering people the chance to explore and learn from person-like interactions. At the same time, these roles ultimately matter because they help us mature and grow into our full human potential, providing both scaffolding and ideals toward which we can aspire.

Work with teaching robots does not show that we are better off without people participating in our normative development. And Zhu et al's work shows that we do not in fact see robots as non-persons, but rather assess them within a personal framework. Instead, robots can help us to build bridges, to connect aspiring people to each other through roles and sometimes-unfamiliar or uncomfortable norms, by letting us look at how these roles function while controlling for complex human variables. Working with and using data from human-robot interaction thus gives us unique opportunities to explore relational ethics accounts and at the same time advance ethical technology development.

Acknowledgements

Many thanks are due: to the excellent and thoughtful organizers and participants of the 2021 *Res Philosophica* conference on the theme "Globalizing Empirical Philosophy", to Jeanine Weekes Schroer, Robert Schroer, Jason Ford, and Laura Engel for their generous discussion and feedback, to the students in my Fall 2021 *Values & Technology* and Spring 2022 *Eastern Philosophy* course for lively engagement with these issues, and to the editor and referee for this journal for their insightful comments, all of which contributed greatly to the development of this project.

References

Nathaniel F. Barrett. 2015. “A Confucian Theory of Shame.” *Sophia* 54 (2): 143-163. https://doi.org/10.1007/s11841-014-0426-0

John M. Doris. 1998. “Persons, Situations, and Virtue Ethics.” *Nous* 32 (4): 504-530.

Nadine Andrea Felber,, Félix Pageau, Athena McLean, and Tenzin Wangmo. 2022. “The Concept of Social Dignity as a Yardstick to Delimit Ethical Use of Robotic Assistance in the Care of Older Persons.” *Medicine, Health Care and Philosophy* 25 (1): 99-110. https://doi.org/10.1007/s11019-021-10054-z

Megan Garber. (2013) “Funerals for Fallen Robots.” *The Atlantic*. September 20. https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallen-robots/279861/ Accessed 6/3/2022

Iris Howley, Takayuki Kanda, Kotaro Hayashi, and Carolyn Rosé. 2014. “Effects of Social Presence and Social Role on Help-Seeking and Learning.” *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*: 415–422. ACM. https://doi.org/10.1145/2559636.2559667

Ryan Blake Jackson and Tom Williams. 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. https://doi.org/10.1109/HRI.2019.8673123

Daniel Leyzberg, Aditi Ramachandran, and Brian Scassellati. 2018. “The Effect of Personalization in Longer-Term Robot Tutoring.” *ACM Transactions on Human-Robot Interaction* (THRI) 7(3): 1–19. https://doi.org/10.1145/3283453

C. Thi Nguyen. 2021. “How Twitter gamifies communication.” In *Applied Epistemology*, edited by Jennifer Lackey, 410-436. Oxford: Oxford University Press.

Nichols, Ryan. 2015. “Civilizing Humans with Shame: How early Confucians altered inherited evolutionary norms through cultural programming to increase social harmony.” *Journal of Cognition and Culture* 15(3-4): 254-284. https://doi.org/10.1163/15685373-12342150

Nuyen, Anh. Tuan. 2009. “Moral Obligations and Moral Motivation in Confucian Role-Based Ethics.” *Dao* 8(1): 1-11. https://doi.org/10/1007/s11712-008-9104-7

Amy Olberding. 2015. “It’s Not Them, It’s You: A case study concerning the exclusion of non-Western philosophy,” *Comparative Philosophy*: 6(2): Article 5. https://doi.org/10.31979/2151-6014(2015).060205

Annie Murphy Paul. 2014. “Robo-Readers Aren’t As Good As Human Readers — They’re Better.” *The Hechinger Report.* April 13.

https://hechingerreport.org/robo-readers-arent-good-human-readers-theyre-better/ Accessed 6/1/2022

Michael Puett. 2015. “Ritual and Ritual Obligations: Perspectives on normativity from classical China.” *The Journal of Value Inquiry*, 49(4): 543-550. https://doi.org/10.1007/s10790-015-9524-7

Bongrae Seok. 2015. “Moral Psychology of Shame in Early Confucian Philosophy.” *Frontiers of Philosophy in China*, 10(1): 21-57. https://doi.org/10.3868/s030-004-015-0003-4

Bryan W. Van Norden. 2002. “The Emotion of Shame and the Virtue of Righteousness in Mencius.” *Dao*, 2(1), 45-77.

David Wong. (2008). "Chinese Ethics", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/ethics-chinese/>. Accessed June 1, 2022.

Qin Zhu, Tom Williams, Blake Jackson, and Ruchen Wen. 2020. “Blame-Laden Moral Rebukes and the Morally Competent Robot: A Confucian ethical perspective.” *Science and Engineering Ethics,* 26 (5): 2511-2526.