The Journal of Sociotechnical Critique

Volume 1 | Issue 1 Article 7

May 2020

“How could you even ask that?”: Moral considerability, uncertainty and vulnerability in social robotics

Alexis Elder
University of Minnesota-Duluth, amelder@d.umn.edu

Recommended Citation
Elder, A. (2020). “How could you even ask that?”: Moral considerability, uncertainty and vulnerability in social robotics. *Journal of Sociotechnical Critique, 1*(1), 1-23. https://doi.org/10.25779/g8a3-f992

“How could you even ask that?”: Moral considerability, uncertainty and vulnerability in social robotics

Cover Page Footnote

A version of this paper was presented at Computer Ethics—Philosophical Enquiry (CEPE) 2019, at Old Dominion University in Norfolk, Virginia, in the United States of America. It was selected to be included in a special issue devoted to papers from the conference, contingent on successful normal blind peer-review processes. Due to disruptions in article revision processes caused by the COVID-19 pandemic, and the very different ways these disruptions have affected different authors, JStC has chosen to publish articles from this special issue as they are ready, rather than waiting to publish them all together. My thanks to the audience at CEPE (Computer Ethics, Professional Enquiry) 2019 for extremely helpful discussion and feedback on an earlier version of this project, as well as Jeanine Schroer, Bob Schroer, Sean Walsh, Laura Engel, Scott Dixon and Thomas Dorr, the students in my senior seminar on robot ethics at the University of Minnesota Duluth in Fall 2018 for the lively conversations that set this paper in motion, and two referees at this journal for their detailed and thoughtful recommendations.

"How could you even ask that?": Moral considerability, uncertainty and vulnerability in social robotics

Alexis Elder
University of Minnesota Duluth

When it comes to social robotics (robots that engage human social responses via "eyes" and other facial features, voice-based natural-language interactions, and even evocative movements), ethicists, particularly in European and North American traditions, are divided over whether and why they might be morally considerable. Some argue that moral considerability is based on internal psychological states like consciousness and sentience, and debate about thresholds of such features sufficient for ethical consideration, a move sometimes criticized for being overly dualistic in its framing of mind versus body. Others, meanwhile, focus on the effects of these robots on human beings, arguing that psychological impact alone can qualify an entity for moral status. What both sides overlook is the importance for ordinary moral reasoning of integrating questions about an entity's "inner life," and its psychological effect on us. Turning to accounts of relationships in virtue ethics, especially those of the Confucian tradition, we find a more nuanced theory that can provide complex guidance on the moral considerability of social robots, including ethical considerations about whether and how to question this to begin with.

Keywords: social robotics, moral considerability, Confucian ethics, moral status, virtue ethics

Asking the right questions

Imagine yourself in a heated discussion with a friend or family member. As the conversation devolves, you each ask more and more pointed questions, until finally, one of you (let's say it's you) challenges the other's standing, and the other person exclaims, "How could you even ask that?"

Depending on how one fills in the details on such a discussion, that last question can look either fair and justified, or unfairly manipulative. That is, sometimes, one person's suspicious behavior, dissembling responses, or otherwise questionable tactics can justify skepticism about, for example, their honesty or integrity. In such cases, questioning the interlocutor's entitlement to inquire about their trustworthiness looks manipulative, or at best evasive, an attempt to change the topic and question the other person's motives rather than submit to further inquiries about their own.

But in other circumstances, it would seem that questioning some things would destroy the very possibility of a conversation. By analogy, consider how Kant pointed out that some elements of cooperative conversation, such as promising, depend on a basic commitment to telling the truth, and without at least some confidence in the other person's honesty, these elements may well prove impossible (Kant, 1998, p. 15). Likewise for questions that may challenge the other person's standing to be taken seriously.

My point here is to draw attention to the fact that sometimes it seems fair to challenge another's standing, and sometimes it does not. Although the connection might not be the most obvious, it seems to me that this question is relevant to robot ethics. In particular, the question about when and why it would be acceptable to interrogate someone's standing in an interaction is germane to issues involving the moral considerability of robots. And the fact that questions that challenge *human* conversational partners' standing are themselves ambiguous is something that I take as an important starting observation.

But before going further, it would be helpful to clarify my starting point in this paper. This is the question: *When and why do robots count as morally considerable?* Variations on this question have been asked for many years, ranging from discussions about moral "status" to current discussions about "robot rights." In this context, I prefer to speak of moral considerability, as this strikes me as relatively broad: "rights" presumes a deontological or legal framework, "status" often invites questions about consciousness (see, for example, summaries of status debates in Gruen, 2017, and LaBossiere, 2017) , but "consideration" can be due for a variety of reasons, including but not limited to such concerns. For this project, then, when I speak of "considerability" I do so to leave open the grounding of this consideration.

I want to remain open in my initial framing of the issue because I think it is important to survey the history of the discussion, in order to see where both plausible and problematic elements can occur. In particular, I think it is helpful to begin by dividing general strategies for answering the considerability question into *internal* and *external* accounts, following Mark Coeckelbergh's lead (Coeckelbergh, 2010). *Internal* accounts focus on internal states of the robot (consciousness, sentience, etc.) as justifying the attribution of moral considerability. *External* accounts focus on external features of the robot, such as appearance, and relational features, such as how its appearance affects human subjects (whether, for example, it provokes empathy, or allows people to practice prosocial behavior), to justify moral consideration. What has been missing, I will argue, is attention to how these concerns fit together. To begin to integrate them, I turn to accounts of interpersonal relationship in Aristotelian and Confucian virtue ethical traditions.

Insides and outsides

Early discussions of ethics in robotics involved questions about what it would take for a robot to be conscious and hence ethically considerable. While robot ethicists rarely try to argue that today's robots count as conscious under any plausible, morally relevant account of consciousness, puzzles related to this issue continue to arise.

One familiar question runs, “When do robots gain moral status?” This then naturally leads to the further question: what, if anything, would be an appropriate test for robot moral status?

Part of the problem is deciding what such a test would measure: common candidates include rationality (LaBossiere, 2017), interests (Neely, 2014), consciousness (Levy 2009), or sentience (MacLennan, 2013), or some other "internal" mental state one takes to be morally relevant. A version of this question also arises in animal ethics, where the question of what internal properties confer moral status is of great interest even when the entities of question seem more likely to have mental lives or subjective experiences than today's robots (Gruen, 2017). This is sometimes taken to involve even further complications that arise when creatures with similar interior lives are widely taken to have different moral statuses—for example, pigs and domestic dogs are both social, emotional creatures with similar cognitive abilities, and yet they occupy very different social and legal positions in many Western countries. This problem might seem less pressing if one is already committed, for example, to ethical vegetarianism—the practical problem of convincing *others* remains, but those persuaded by, for example, Peter Singer’s or Tom Regan’s accounts of animal ethics will already be committed to the view that to distinguish between puppies and piglets is morally unjustifiable (Norcross, 2004; Regan, 2004; Singer, 2009). This does, however, introduce questions about what might justify treating psychologically similar creatures differently (Coeckelbergh and Gunkel, 2014, Darling, 2016).

The question is further complicated by issues about what kind of external observables could count as plausible evidence of mental states, a deep problem in philosophy of mind. In the field of artificial intelligence, the Turing Test is sometimes taken to give at least defeasible criteria for mentality, but its status is highly controversial, and Turing may have taken the term artificial *intelligence* to itself be an important means of deflecting any ethically significant implications being attached to a mechanism that passes the test, although this has been debated (Turing, 2004; Estrada, 2018). The dual uncertainties of what we ought to measure, and how we ought to measure it, lead some to argue that an appropriate ethics for our treatment of robots ought not presume a perfect measuring tool but should instead take into account the relative risks of getting an entity's moral status wrong in both directions: false positives and false negatives (LaBossiere, 2017; Neely, 2014).

Here, speculative robot ethics intersects with practical concerns of today's roboticists, as the costs and benefits associated with false positives are of immediate interest. Research in HRI (human-robot interaction) shows that people—including informed, technical users—attribute agency, intentionality, and moral significance to even simple robots, based on cues like apparently-autonomous movement and the appearance of stylized faces (Darling, 2012; Scheutz, 2012), and they seem to work by appealing to non-rational but influential portions of human psychologies, ones that affect our decision-making and altruistic tendencies (Bateson, Nettle, & Roberts, 2006; Farah & Heberlein, 2007; Haley & Fessler, 2005).

Risks of these “false positives” include vulnerability to manipulation by robot manufacturers (Scheutz, 2012) and hackers (Grodzinsky, Miller, & Wolf, 2015), a tendency to decrease appreciation for genuine agents in exchange for a placid acceptance of "real enough" (Turkle, 2011), substitution of robot companions for human companionship (Sharkey, 2014), and a tendency to displace blame from human operators and manufacturers to the robots themselves, thus allowing human agents to duck responsibility illegitimately (Bryson, 2010; Sharkey, 2017; Vallor & Bekey, 2017). These concerns tend to support the idea that we are justified in questioning whether they ought to be treated as morally considerable.

However, against conservatism about moral considerability, a number of considerations make it plausible that these “external” features of robots might give grounds to take them to be morally considerable, all on their own. For example, it seems problematic to interrogate real-seeming person-ish creatures and demand that they prove themselves to have moral status; this is not an ordinary feature of moral experience and in fact seems itself somewhat unethical (Coeckelbergh, 2010; Coeckelbergh & Gunkel, 2014; Gunkel, 2018). That is, it seems appropriate to criticize someone for questioning apparent moral status. Conceptualizing robots as subservient, and “dehumanizing” them, can manifest troubling tendencies toward oppression and domination (Estrada ,2020). Furthermore, mechanisms that contribute to human tendencies to attribute moral significance to robots may be valuable and worth protecting, even when they “misfire” (Darling, 2016, 2017), and sociable robots may be valuable in part when and because they help us to rehearse and reinforce prosocial tendencies among human beings (Sullins, 2008), similar to some psychological benefits of companion animals. These external approaches, sometimes dubbed “relational” accounts because they focus on (external) relations between the entity and human beings with whom it interacts (Coeckelbergh & Gunkel, 2014), challenge the assumption that the correct way to assess ethical issues involving non-human entities is to try to determine which (internal) properties of an entity are necessary and sufficient for moral status, in order to detect or intuit with some degree of certainty whether any given entity does so. Instead, they focus on moral grounds for taking seriously those observable features that, with some regularity, engage morally significant responses of our own.

Missing pieces

There are risks no matter which strategy we adopt. It looks like the options are to make ourselves vulnerable to manipulation or to impose unreasonably high standards for moral consideration that rule out many genuinely deserving subjects and impair our potential to engage in relationships. Furthermore, it is unclear how these concerns ought to be balanced, owing to contested issues that span across a variety of fields in philosophy, from the moral significance of empathy and emotional reactions (Darling, 2012; Coeckelbergh, 2018b) to the plausibility of philosophical "zombies" (Gray and Wegner, 2012). What is missing in internal-properties accounts, as some have argued (Coeckelbergh and Gunkel, 2014) is respect for the moral importance of relationships, and what is missing in relational accounts is an appreciation for the importance of truth-conditions for beliefs about those things with which we take ourselves to be related (Nyholm, 2020). Consider the example with which this paper began: concerns about manipulation and vulnerability are themselves to be found within relationships even among uncontroversial instances of persons.

Both "sides," as I have articulated them here, overlook something important about the other: the "internal qualities" theorist overlooks, among other things, the fact that apparent possession of sentience or consciousness has not resulted in anything like widespread or consistent ethical protections for non-human animals. Internal qualities, if they matter, call for a widespread overhaul of current ethical practices, and do not uncontroversially support ethical status in and of themselves. But more importantly, they overlook the importance of relations and relationships: that even a widely-recognized-as-a-person's moral significance can vary to another person based on whether that someone is a friend, a relative, or a stranger—while we might owe basic respect to all those who qualify as people, someone's being *my* child can make a significant difference in how I ought to spend my time and energy. And furthermore, that even clear non-persons can have moral significance based on other factors: work of arts, objects of sentimental attachment (to oneself or others), religious artifacts (even of someone else's religion), and so on.

Meanwhile the “relational” approaches miss the opposite point: that relationships are not just about how *others* make *us* feel, but about how people interact and respond to each other; become emotionally interdependent, and open to each other’s perspectives on the world. (This is of course not true of all relationships—enemies, rivals and so on—but is characteristic of valuable interpersonal relationships, and to the extent that other relationships lack this, they seem to be morally problematic.) Consider, for example, Freya Mathews’ discussion of why relationships with “pet” animals can be distinctively valuable:

> …emotional involvement with creatures who do not share our human goals and aspirations, our systems of values, enables us to gain an external perspective on those values. It enables us to appreciate how odd or arbitrary our human priorities might appear to non-human observers (Mathews, 2007, p. 16).

While something can present itself to us as person-like or even friend-like via fairly simple visual cues, such as eyespots or particular movement patterns, like the famous Heider and Simmel animation of two triangles and a circle (Schroer & Schroer, 2019), moving from “activates semiautonomous person-detector neural system” to a real relationship requires opening oneself to another’s point of view and concerns, which is difficult if the other has no point of view or concerns.

Rather than think we need to choose between an “internal” and an “external” approach, I think we ought to consider how to integrate them. Instead of a purely “relational” account, we would do well to start with a richer account of relationships – including both relations and relata – may be able to reframe concerns about social robots in light of skepticism, risk, and vulnerability in relationships. We need to think about when it is appropriate, and when it would be inappropriate, to challenge their status versus go along with appearances, and we can take advantage of existing discussions about appearance and emotion in interpersonal relationship to do so.

Reasoning in and about relationships

In thinking through responsible management of the inevitable risks posed by the appearance of relationship, one can take account of the problems posed by robots and reinforced by debates in animal ethics. Such an account can shed light on both speculative and immediate concerns of robot ethics. It can articulate ethically appropriate means of investigating appearances, within the limits of our epistemic abilities.

That this is so is perhaps best demonstrated by considering the roles of *deception*, *trust*, and *honesty* in interpersonal relationships.

These things matter in interpersonal relationships because we care about more than appearances; more than *how* someone makes us feel or *whether* something is a person. We care about the extent to which someone's appearances match their inner states. Thus, some so-called "relational" externalist accounts miss the mark in thinking that skepticism about the genuineness of a creature's appearance is incompatible with thinking of them as people, or that approaching personish things in a relational way requires accepting things at face value.

This manifests in different ways in different traditions. For example, in the *Nicomachean Ethics*, Aristotle has the following to say about deception and friendship:

> one might complain of another if, when he loved us for our usefulness or pleasantness, he pretended to love us for our character. For, as we said at the outset, most differences arise between friends when they are not friends in the spirit in which they think they are. So when a man has deceived himself and has thought he was being loved for his character, when the other person was doing nothing of the kind, he must blame himself; when he has been deceived by the pretences of the other person, it is just that he should complain against his deceiver; he will complain with more justice than one does against people who counterfeit the currency, inasmuch as the wrongdoing is concerned with something more valuable. (Aristotle, 1999, p. 140/1165b5)

This focuses on the experience of being deceived and is both critical of a tendency to self-deception and the deceptiveness of others. And while Platonic philosophy focuses on metaphysical concerns and the appearance/reality gap (as for example in the allegory of the cave in his *Republic*), Aristotle endorses a commitment to starting from and reasoning through appearances, preserving them as insofar as possible (Aristotle 1999, p. 100/1145b2-7) He is arguing within the scope of the appearances that people's behavior can generate expectations that are ill-matched to the situation, and that thriving as a person whose wellbeing requires relationship (p. 119/1155a5-10) requires the ability to situate judgments based on (temporary) appearances within broader and longer-standing contexts of reliability and mutual expectations. That is, he can be taken to be concerned with aligning one's judgments about internal and external states, without presuming epistemic access to a reality "behind the appearances" altogether. We have to grapple with our own (sometimes involuntary) tendencies to interpret ambiguous data in our favor when engaging in social relationships, as well as others' ability to mislead us for their own ends.

Confucian virtue ethics integrates a concern for the social roles we occupy into its account of the virtues, which informs how Confucius approaches the issue of honesty in interpersonal relationships. He speaks to the importance of being *xin* if one is to be a virtuous individual in a highly social context. What is *xin?* As Wee argues, although Confucian translators usually render it as “trust,” *xin* is really more about one's deeds matching one's words, out of a sincere commitment to being transparent to others (Wee, 2011). Confucius is clear that this is not meant to imply that the person is good overall, since a small-minded stickler can be *xin*, but it is an important step on the way to being a virtuous individual. *Xin*, Confucius says, will make one’s interactions go better; it is difficult to achieve, but living up to one’s word conscientiously is a necessary prerequisite for interpersonal harmony.

In the *Analects* (15.6), a conversation between Confucius and his student makes this point especially clear.

> Zizhang asked about getting by in the world.
> The Master [Confucius] replied, 'In your speech, be dutiful and trustworthy, and in your conduct be sincere and respectful. In this way, you will always get by in the world, even if you find yourself in some barbarian state. If your words are not dutiful and trustworthy, and your conduct is not sincere and respectful, how can you possibly get along, even in your own region? When standing still, visualize these principles standing by your side; when riding in your carriage, see them resting before you on the crossbar. Only then will you get by in the world.”
> Zizhang then wrote these words on the end of his sash. (Confucius, 2003, p. 176/Analect 15.6)

This passage helps make clear why Wee differentiates it from the English word “trust:” *xin* is not the same thing as interpersonal trust, since *xin* is about the individual rather than the relationship. One is *xin* even amongst the barbarians. And yet, it is important as a part of the foundation that makes interpersonal relationships possible.

Whether focused on self-deception and correct interpretation of others’ intentions, or holding oneself to a standard for others, both Aristotelian and Confucian virtue-ethical traditions, then, emphasize the importance of accurately discriminating between the mere appearance of friendliness and the genuine friend, a concern that informs both how one chooses to invest in interpersonal relationships with others, as Aristotle argues, and how one conducts oneself in social settings, as Confucius points out. Both philosophers hold that morally valuable relationships involve emotions, but consist of much more than mere gut-level reactions to eyespots and movement.

Both Aristotelian and Confucian virtue ethics also praise the value of relationships but caution against deceptive appearances. It would seem, then, that both traditions would be critical of what I have characterized as “false positives,” mistaking the superficial appearances of social robots that trigger human social responses for the real presence of people worthy of relationship.

However, it does not follow from this that we ought to avoid empathetic or social responses to these cues, either. Not every engagement of emotional or social responses that is triggered by a non-person entity should be taken to be a “false positive.” And on this topic, Confucian ethics offers advantages over Aristotelian for its sophisticated moral psychology and discussion of the social conditions and practices that promote virtue. In particular, Confucian ethics’ emphasis on *li*, sometimes translated as “rituals” or “etiquette” but perhaps better understood for our purposes as structured social practices that facilitate character development and flourishing relationships, is of use when thinking about how and when we ought to take robots to be morally considerable, at least in the absence of clear reason to attribute “internal” states such as consciousness or sentience.

Confucian resources on emotion-stimulation

Specifically, Confucius and his followers offer detailed discussions of music and funerals that can be helpful in thinking about social robots and moral consideration. Music, because it triggers powerful emotional responses, and funerals, as extensive social practices developed (on Confucian accounts) in order to show respect to corpses, which trigger strong emotional responses for the bereaved by reason of their resemblance to the person that was, and yet which are not conscious or sentient and (in the naturalistic worldview common among the Confucians) offer no prospects of becoming so.

Music as tool for cultivating emotion

Confucius praises music for its ability to stimulate emotional responses and so sensitize people to feel emotions appropriate to the situation. He is critical of some forms of music for emphasizing what he takes to be the wrong sorts of responses, preferring, for example, ancient Shao and Wu music as opposed to the more seductive contemporary Zheng music (E.G. Slingerland, 2014, p. 73). But he is optimistic about music’s potential for moral education. At the same time, he does not think one ought to rely on it to do all of the moral work; when visiting a town in which the administrator has encouraged all of the citizens to play music, apparently to excess, Confucius remarks that he is using “an ox-cleaver to kill a chicken”—a memorable phrase to call out overkill (Confucius, 2003, p. 200/Analect 17.4)

Confucian scholar Mencius follows up on this interest in music as a tool for moral education, remarking:

> The content of benevolence is the serving of one’s parents; the content of dutifulness is obedience to one’s elder brothers; the content of wisdom is to

> understand these two and to hold fast to them; the content of the rites [*li*] is the regulation and adornment of them; the content of music is the joy that comes from delighting in them. When joy arises how can one can stop it? And when one cannot stop it, then one begins to dance with one's feet and wave one's arms without knowing it. (Mengzi, 2004, p. 87/4A.27)

We can imagine, then, that a similar defense might be made of social robots: Without even knowing it, one reacts with empathy or social engagement, and this can be incorporated via *li* into a kind of regulated adornment of our natural tendencies toward empathy and sociality, in order to develop into better human beings.

Social robots and music are similar in that they both can engage our emotional responses involuntarily and unreflectively. While in robot ethics this is taken to raise worries about manipulation, or else treated as valuable in its own right (as when one treats activation of empathy as good in itself), by considering how music can involuntarily engage our emotions in a way that can seem valuable for its own sake; moving with the music becomes an end in itself. And while music can certainly be used to stir emotions in the service of other ends, including morally problematic ends, it would be a mistake to thereby treat it with suspicion simply because it is not under our voluntary control.

Against worries about manipulation, in fact, the Confucian can point to music's tendency to involuntarily draw us in and set us dancing as a potentially valuable technology to provoke emotional responses as part of the project of developing desirable human traits. It will not do so all on its own and too much can be overkill (think of the ox-cleaver!) but can be a valuable part of a morally enriching life. To the extent that music is morally valuable, so too can social robots be. And to the extent that we think music can be morally considerable; worth considering and worth protecting, without getting into debates about whether it is conscious or sentient (think about how silly that would be!) we might take robots to similarly be candidates for a parallel form of moral consideration without attributing to them "internal" qualities like consciousness.

Funerals as tools for cultivating social dispositions

Another topic in Confucian ethics that bears on moral considerability of social robots, is, as I have said, discussion of funerals. It is important to note that Confucius and his followers were agnostic, at best, about an afterlife, and their arguments framed strictly in terms of life on earth. At one point in the *Analects*, his pupil Zilu apparently "asked about serving ghosts and spirits. The master [Confucius] said, 'You are not yet able to serve people—how could you be able to serve ghosts and spirits?' 'May I inquire about death?' 'You do not yet understand life—how could you possibly understand death?'" (Confucius, 2003, p. 115/Analect 11.12) But even without understanding death, sooner or later we all must deal with it—and not just our own, but the deaths of people around us. Chinese ethics of the time involved extensive debates about the value and appropriateness of funeral practices, and Confucian scholars articulated a defense of this form of *li* that was grounded in the practical and psychological benefits to human beings.

One particularly prominent defense comes from Confucian scholar Xunzi, in his essay “On Ritual.” In her paper “From corpses to courtesy: Xunzi’s defense of etiquette,” which focuses on this work, Amy Olberding identifies a number of ways that his concerns may be incorporated into modern ethics.

In contrast to a Kantian approach that emphasizes respect for persons, Olberding explains,

> Xunzi’s defense of etiquette [*li*] issues from a more prosaic sensibility: the joint recognition that cooperative, communal life requires that we find human company appealing and attractive, and that human beings in their natural state are, if not as a rule then too often for our shared good, unappealing and unattractive to their fellows (Olberding, 2015, p. 147).

Furthermore, it avoids emphasizing our “natural” or “authentic” impulses, instead arguing that its value comes from its artificiality. While human nature left to its own devices can sour us on each other, rituals (broadly construed to include both funerals and etiquette) can help us to cultivate dispositions that counter these tendencies while facilitating social cooperation.

Olberding begins with Xunzi’s discussion of funeral rituals to illustrate the need for artificial intervention. Funerals are “justified by the corpse’s ‘natural transformation’—that is, its decay—and our responses to it: Ritually adorning the corpse is a strategy that ‘disguises its hideousness.’” (Olberding, 2015, p. 150) While eyespots and movement may trigger our innate social responses, which as we have seen is a concern for robot ethics, corpses trigger innate feelings of disgust, which counts as a concern for funeral ethics. What they have in common is that corpses strongly resemble human beings, in ways that, like robots, bring up social responses, and responses we do not want to suppress. Olberding explains,

> Where under most conditions we would simply distance ourselves from what disgusts us, the corpse does not merely disgust but is simultaneously identified as the person we grieve. It is at once thing and person, a thing we must get rid of and a person we long to keep but cannot. Corpses are disposal problems with special features, for the bereaved will neither readily nor easily disidentify the corpse with the living person it symbolically represents. (Olberding, 2015, p. 151)

That is, it seems *good* to identify the corpse with the person lost, even if one must also face the reality that the person is no longer there. This is consistent with the externalist point that it can be good to react to social robots with empathy, even while agreeing with internalists that these robots lack important internal features, situating both within our experience as creatures that care for others by way of emotional responsiveness to appearances while recognizing that appearances are not all that matters.

She explains that on Xunzi's account, funeral rituals create "conditions under which corrosive disgust that would undermine affection and respect cannot find purchase, and ethically fruitful longings and attachments are affirmed and encouraged." (Olberding, 2015, p. 152) This is done by using rituals to shape our emotional responses and habituate us into responding to other human beings in more cooperative and considerate ways.

At this point, one might be forgiven for thinking that the thrust of Xunzi's argument focuses on the individual impact of corpses (or robots) on human psychologies, such that it might end up being more appropriate to embrace his approach with some people than others. For example, many parents have actively solicited "politeness modes" for artificial personal assistants like Alexa and Siri, in order to help their children practice good manners and counteract these devices' tendencies to encourage them to bark requests at people (Metz, 2018). But perhaps adults are less suggestible, perfectly capable of partitioning off their treatment of robots from their treatment of human beings, rendering ritualized shaping of emotional responses superfluous or paternalistic. But in addition to taking adults to be capable of struggling with the pull of emotion in the presence of a corpse even when they understand that the person is no longer there, in ways that funeral rituals can help with, Xunzi notes that people's treatment of others can itself have an influence on others (Xunzi, 2014, p. 257/380-390), and thus we need not only funeral rituals for corpses but also etiquette to govern our interactions with each other, because the rude person does not only influence the person to whom they are rude, but anyone who happens to observe their interaction (Olberding, 2015, p. 156).

Olberding points out that this is "consonant with evidence in empirical psychology. ...The phenomenon of "emotional contagion." Thus, "we profit where our practices manage and regulate the atmospherics of shared space and social encounter to favor affirmations of fellow feeling and humanity." (Olberding, 2015, p. 156) While often criticized for their artificiality, funerals and etiquette, she says, are "clever strategies for protection and flourishing" and "'artificial' much the way a house is… shelter from natural elements that would imperil our well-being and, at their more ambitious, serving our hopes to find ourselves at home with others in lives that are shared."(Olberding, 2015, p. 159)

This highly pragmatic account of the value of emotionally engaging social robots might sound merely instrumental (we value robots because they help us cultivate important prosocial dispositions) but this justification is not appropriate as a *motivator* of our engagement with them, precisely because purely instrumentalist interactions with people are themselves failures of prosocial attitudes. To think of something as a mere tool for one's use (even to practice prosocial dispositions) would tend to undercut its effectiveness, when the disposition it is used to cultivate is one that of necessity involves direct consideration for and responsiveness to persons. The ways that this complicates the picture involve both how we view our relationship to our emotional responses, and how we see ourselves as relating to others, as the next two sections argue, and ultimately explain why it can be so difficult in practice to draw clear lines dictating when it is appropriate to challenge something's moral standing.

Robots as cultivators of social dispositions

This same line of reasoning we saw with music can be extended to our treatment of robots: to treat them as morally considerable need not facilitate "pernicious falsity" but rather an (admittedly artificially induced) way to cultivate our social selves, both by direct means (as for example by encouraging us to say "please" and "thank you" and to treat those things that stimulate altruistic feelings as worthy of attention—because these feelings are not a scarce resource but dispositions to be extended through practice) and because of the impact our own treatment of robotics may have on others. *Slate* columnist Rachel Withers recently authored an essay titled "I don't date men who yell at Alexa," in which she argues that verbal abuse of virtual assistants can reveal dispositions and biases that reflect gender and power imbalances. Her titular decision seems eminently reasonable to me, because the gendered character of personal assistants such as Alexa invoke further overtones of misogyny that can be off-putting to others, an area where it is reasonable to maintain this distaste for such behavior when we already have an uphill battle to ensure that people grant women the moral consideration they are clearly owed (Withers, 2018).

At the heart of both the Confucian theories articulated above, and contemporary debates about the impact of social robotic technologies on human beings, is a question about how we relate to our sometimes-involuntary emotional responses, and, depending on how that question is answered, what we ought to do about this.

In a discussion of historical trends in European thought, Mark Coeckelbergh characterizes a divide that emerged between Romanticism and Enlightenment rationality (Coeckelbergh, 2018). The Enlightenment rationalist says emotions distort and misrepresent reality, and so the perfect reasoner should do away with them. This approach is well-captured in a line from Arthur Conan Doyle's story "A scandal in Bohemia," in which the protagonist Sherlock Holmes's relationship to emotions and reasoning is described:

> All emotions... were abhorrent to his cold, precise but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen, but as a lover he would have placed himself in a false position. He never spoke of the softer passions, save with a gibe and a sneer. They were admirable things for the observer—excellent for drawing the veil from men's motives and actions. But for the trained reasoner to admit such intrusions into his own delicate and finely adjusted temperament was to introduce a distracting factor which might throw a doubt upon all his mental results. Grit in a sensitive instrument, or a crack in one of his own high-power lenses, would not be more disturbing than a strong emotion in a nature such as his. (Doyle, 2019, p. 163)

By contrast, the Romantics reacted against this framing by paradigmatically holding that emotions are valuable in themselves, regardless of how well they track what's going on in the world, and thus present 'reasons' for action that outweigh what 'reason' would have us do.

But these extremes that some thinkers have inherited from the division are not the only options. One could, for example, take the view that emotions should be observed, and can be useful data points, but are not to be taken as definitive in their own right.[1] One might think that emotions are like small, excitable dogs—when they sound an alert, you should investigate and see what they're responding to, but exercise detached judgment about how to ultimately respond, because they can be triggered by both serious and insignificant stimuli. In a therapeutic context, this can seem better than either validating every emotional response as worthy of the action the emotion recommends, or trying to clamp down on emotions via sheer force of will or denial, but ultimately is aimed at working with whatever emotions arise, rather than adopting a normative attitude toward emotions—that is, toward evaluating whether or not one ought to feel a particular way, or what one should do about an emotion that seems in some sense inappropriate.

But the Confucian view, which seems to me more nuanced, holds that emotions are important parts of mature reasoning, but that people need emotional training in order to be reliable and to reliably direct attention to what matters. This training occurs through rituals, through music, through poetry and stories, and is described via gardening metaphors, especially by

[1] This is a view often associated with mindfulness practices—I refrain from calling it "Buddhist" as this would oversimplify the complexities of actual Buddhist theories of cognition and introspection, see Purser, 2019; for a summary of actual Buddhist philosophy of mind see Coseru, 2017.

Mencius (Mengzi, 2004, pp. 2A:6, 6A:6). We start off with promising sprouts, but they need nurturing and sometimes pruning as well as fertilizing to grow into dependable parts of the mature person. It is this last framework that I think is most promising for dealing with social robotics: many of the emotional and social responses they stimulate in us are important and worth cultivating, but also need refinement if they are to serve us well in complex social situations.

We ought not treat our feelings about robots as definitive of their value, nor do we want to be suspicious of them merely because they provoke emotional responses and thereby "interfere with reasoning." We could take a kind of mindful, detached, playful or ironic view, one that leads to a kind of one-step-removed emotion-observing, but we can also treat them as potential cultivators of worthy emotional responses (and therefore also potentially abusable), following the Confucian framework—they would be like other rituals or music or stories. To respond to people with appropriate seriousness means we need to think about the emotional engagement that this involves, and how to cultivate it, including when not to question it.

Role ethics and robotics

Against concerns that these might still inflate the importance of emotional responses over the facts about personhood (or lack thereof) attributable to robots, Confucianism offers another advantage. The sophisticated *role ethics* developed in Confucian philosophy can be useful in thinking about the ethical consideration due a role versus a person, which can be helpful when the robot occupies a role also filled by human beings, like that of caregiver, companion, teacher, or housemate.

Confucianism famously involves role ethics[2], the idea that in virtue of the social roles one occupies, one thereby acquires ethical reason to act in particular ways. These social roles can include professional roles, such as teacher or student, but also family roles, like parent or child. In particular, Confucianism is often associated with *filial piety*, which emphasizes responsibilities associated with being the child of one's parents, and in particular the importance of loving regard, gratitude, and care for one's parents. While one might be tempted to think of practicing piety as instrumentally valuable in the service of caring for the intrinsically valuable people who occupy this role, for the Confucian, the roles themselves can be directly valuable, even when the people who fulfill them are not in themselves deserving of the value assigned to the role. Filial piety can occur naturally, as it were, when parent-child relationships are harmonious and unproblematic, but not every parent-child relationship is smooth. In

[2] For a survey of Confucian role ethics, see Ramsey, 2016.

particular, filial piety can be challenging because not every parent is a good parent. Confucians recognized this, and offered detailed accounts of how filial piety can be exercised when parents are problematic. For example, the sage-king Shun was a moral exemplar, but also the child of terrible parents, who (among other things) repeatedly tried to kill him. Confucian discussions of Shun's relation to his parents included obligations to protect himself from them (Huang, 2012, p. 43) and in some cases circumvention of the formal practices of filial piety, in order to protect what remained of his emotional regard for them. David Wong notes, for example, the story of his marriage as an example of this latter issue. Shun married without his parents' approval, a violation of the customs of filial piety. When challenged on this, he explained that if he had gone to them for approval, they would have denied it simply to hurt him, and then he would have been obligated to obey, and would have naturally resented them. Instead, in order to prevent this harm of resentment, he married without their permission, thus preserving what he could of his relationship with them. (Wong, 2011, p. 8)

While one might not agree with every one of the details of the discussion, the core idea that we sometimes have to work with people's limitations in order to preserve what is valuable about our relationships, and the roles we acquire in virtue of them, may be familiar to many of us. Because Confucian role ethics provides a framework for thinking about how to relate to roles versus people, as Shun related to his parents as vicious people who nevertheless occupied important roles for him, it can offer conceptual resources for thinking about the roles robots occupy in relation to us, while recognizing their limitations in filling these roles. Even if it did not matter to Shun's parents as individuals that he thrive, in virtue of his relationship to them as their child, it was important to *him* that his parents not harm him or get in the way of his flourishing—over and above his general concern that others not so harm or interfere with his happiness. And this can be a subtle but important point, one perhaps familiar to many children with difficult parents. Even when actual parental abilities are lacking, recognizing the importance to oneself as child that one's parents can have is an important part of coping with the limitations of these people in one's life.

One thing that role ethics makes clear is that our relationships to roles can matter, over and above relationships to people, and this is consistent with a point that is often overlooked in robot ethics: We err when we assume that harming a sentient or conscious being is a *necessary* rather than *sufficient* condition for wrongdoing. It would surely be problematic, for example, to abuse robots that turn out to be sentient. But it is not necessary that they experience the harm in order for it to be wrong for us to treat them in particular ways, any more than a corpse needs to feel pain in order for it to be wrong for us to mutilate it, or (in Xunzi's very practical example) a small child need not understand that it is being humiliated in order for it to be wrong for us to humiliate them—and we need not ignore robots' non-person-ness in order for them to matter to us. And, again, this is not a matter of merely instrumentally performing respect for children or corpses as practice for the "real" cases; it matters for how we think of each other to treat others as worthy of care even when they are not in a position to complain, which means we ought to take them to be intrinsically valuable in our lived experiences.

Conclusion

Returning to the example with which this paper opened: challenging appearances in interpersonal relationship is a tricky business. It can be entirely appropriate to reject attempts to question them, as the titular question runs: "how could you even ask that?" Some disposition to take others at face value is necessary for interpersonal relationships to be possible, and while current robots' "faces" may fail to represent internal conditions necessary to ground nuanced and sophisticated human interpersonal relationships, we should hold ourselves to standards that make room for others to fill social roles. At the same time, we may need to be alert to the possibility that appearances and role occupancy can both be imperfect indicators of ability, and train and retain the skillset necessary to navigate deception and disappointment—in interpersonal human-human relationships as well as human-robot ones. A nuanced account of respect for role can make room for healthy skepticism about potentially misleading or accidentally misleading appearances, without licensing abusive or disrespectful treatment of the entities that occupy social roles for us, while endorsing careful attention to both benefits and risks of emotion-eliciting technologies, so as to develop ethically important emotional dispositions. One can respect the role a social robot occupies, without falsely believing its superficial appearances make it a "person" in all of the ways that might be relevant to philosophers—or to us. This respect can give us reason to avoid abusing or disrespecting these robots, and we can give thought to the rituals it would be appropriate to employ around them, given their effect on our social responses. The emotions they inspire, from sociality to scorn, can be worthy of consideration as part of our concern for our own ethical development, and like music or funerals, they can be constructed and used so as to amplify cooperative and pro-social emotional dispositions and discourage those emotional responses that interfere with our ability to live well with others—like (potentially) frustration, scorn, arrogance, gullibility or superiority.

For example, take the robot funerals that Julie Carpenter has documented in the U.S. military (Garber, 2013).[3] Within a Confucian framework, the rationale for such a ceremony might turn out to be remarkably similar to

[3] I thank an anonymous reviewer at this journal for suggesting the relevance of this case.

that Xunzi offers for human funerals. While it makes no difference to the guest of honor, it marks a way for people to honor the significant role the deceased has played in their lives, and to give their emotion and attachments space to be felt and recognized without "running wild." The roles that life-saving bomb detection and pack robots play in these soldiers' lives can be appropriately recognized as valuable and, in virtue of their connections to human soldiers, responses to their loss can be given weight in the form of ritual, as part of protecting the social capacities and human lives in which both humans and robots are embedded.

Thus, robots can turn out to be morally considerable—that is, worthy of moral consideration—without reducing the value of relationships to emotional reactivity on our part, as the externalists sometimes suggest, and leave room for empirical research on, for example, the psychological mechanisms and effects of empathy, without waiting for robots to become sentient or conscious, or for us to work out reliable means of determining that they have done so. Here, the Confucian idea of *li* as ritual or etiquette can be especially helpful because it avoids the universality often associated with moral norms, emphasizing the importance of contingent social practices within a particular place and culture, which nonetheless give us a kind of structure for shaping our social selves in response to the challenges we tend to face within some shared context. By reflecting on what would count as justifiable rituals for us to practice, and thinking about how particular reactions accord with other social concerns, we gain tools to help us navigate when it is appropriate to accept them at face value, and when to question appearances.

References

Aristotle. (1999). *Nicomachean ethics* (2nd ed.; T. Irwin, Trans.). Indianapolis: Hackett Publishing.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412–414. https://doi.org/10.1098/rsbl.2006.0509

Bryson, J. J. (2010). Robots should be slaves. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, 63–74.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, *12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5

Coeckelbergh, M. (2018a). *Environmental skill: Motivation, knowledge, and the possibility of a non-romantic environmental ethics.* S.l.: Routledge.

Coeckelbergh, M. (2018b). Why care about robots? Empathy, moral standing, and the language of suffering. *Kairos: Journal of Philosophy & Science, 20*(1), 141-158.

Coeckelbergh, M., & Gunkel, D. J. (2014). Facing animals: A relational, other-oriented approach to moral standing. *Journal of Agricultural and Environmental Ethics*, *27*(5), 715–733. https://doi.org/10.1007/s10806-013-9486-3

Confucius. (2003). *Analects: With selections from traditional commentaries.* (E.G. Slingerland, Trans.). Retrieved from https://books.google.com/books?id=6DseYHSfaagC

Coseru, C. (2012, October 12). *Mind in Indian Buddhist philosophy.* The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2017/entries/mind-indian-buddhism/

Darling, K. (2016). "Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects". In *Robot Law* (pp. 213-232). Cheltenham, UK: Edward Elgar Publishing. https://doi.org/10.4337/9781783476732.00017

Darling, K. (2017). “Who’s Johnny?” Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0: From autonomous cars to artificial intelligence* (pp. 173–188). Oxford: Oxford University Press.

Doyle, A. C. (2019). *Complete Sherlock Holmes*. S.l.: Chartwell Books.

Estrada, D. (2018, December). Value alignment, fair play, and the rights of service robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 102-107).

Estrada, D. (2020). Human supremacy as posthuman risk. *Journal of Sociotechnical Critique, 1*(1), 1–40. Advance online publication. https://doi.org/10.25779/j5ps-dy87

Farah, M. J., & Heberlein, A. S. (2007). Personhood and neuroscience: Naturalizing or nihilating?. *The American Journal of Bioethics*, *7*(1), 37-48.

Garber, M. (2013). Funerals for fallen robots. (September 20). *The Atlantic.* Retrieved August 24, 2020, from https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallen-robots/279861/

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125-130.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2015). Developing automated deceptions and the impact on trust. *Philosophy & Technology*, *28*(1), 91–105. https://doi.org/10.1007/s13347-014-0158-7

Gruen, L. (2017, August 23). *The moral status of animals*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2017/entries/moral-animal/

Gunkel, D. J. (2018). The other question: Can and should robots have rights? *Ethics and Information Technology*, *20*(2), 87–99. https://doi.org/10.1007/s10676-017-9442-4

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? *Evolution and Human Behavior*, *26*(3), 245–256. https://doi.org/10.1016/j.evolhumbehav.2005.01.002

Huang, Y. (2012). *Confucius: A guide for the perplexed*. New York: Continuum.

Kant, Immanuel. (1998). *Groundwork of the metaphysics of morals* (M. Gregor, Trans.) Cambridge: Cambridge University Press.

LaBossiere, M. (2017). Testing the moral status of artificial beings; Or "I'm going to ask you some questions..." In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 293–306). Oxford: Oxford University Press.

Levy, D. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics, 1*(3), 209-216.

MacLennan, B. (2013). Cruelty to robots? The hard problem of robot suffering. In *Proceedings of the 2013 Meeting of the International Association for Computing and Philosophy (IACAP*).

Mathews, F. (2007). Without animals life is not worth living. *Between the Species*, *13*(7), 1–28.

Mengzi. (2004). *Mencius* (Revised edition; D. C. Lau, Trans.). London: Penguin.

Metz, Rachel. (2018). Please, Alexa? Amazon’s new parental controls will encourage politeness. (April 25). *Technology Review.* Retrieved August 24, 2020, from https://www.technologyreview.com/f/611019/please-alexa-amazons-new-parental-controls-will-encourage-politeness/

Neely, E. L. (2014). Machines and the moral community. *Philosophy & Technology*, *27*(1), 97-111.

Norcross, A. (2004). Puppies, pigs, and people: Eating meat and marginal cases. *Philosophical Perspectives*, *18*(1), 229–245. https://doi.org/10.1111/j.1520-8583.2004.00027.x

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism.* Lanham, MD: Rowman & Littlefield Publishers.

Olberding, A. (2015). From corpses to courtesy: Xunzi’s defense of etiquette. *The Journal of Value Inquiry*, *49*(1–2), 145–159. https://doi.org/10.1007/s10790-014-9466-5

Plato. (2004). *The Republic*. (C.D.C. Reeve, Trans.) Indianapolis, IN: Hackett.

Purser, R. (2019). *McMindfulness: How mindfulness became the new capitalist spirituality.* Watkins Media Limited.

Ramsey, J. (2016) Confucian role ethics: A critical survey. *Philosophy Compass, 11*(5), 235–245. https://doi.org/10.1111/phc3.12324

Regan, T. (2004). *The case for animal rights* (Updated with a new preface, [2004 ed.]). Berkeley: University of California Press.

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, Massachusetts: MIT Press.

Schroer, R., & Schroer, J. W. (2019). Putting psychology before metaphysics in moral responsibility: Reactive attitudes and a “gut feeling" that can trigger and justify them. *Philosophical Psychology*, *32*(3), 357–387.

Sharkey, A. (2014). Robots and human dignity: A consideration of the effects of robot care on the dignity of older people. *Ethics and Information Technology*, *16*(1), 63–75. https://doi.org/10.1007/s10676-014-9338-5

Sharkey, A. (2017). Can we program or train robots to be good? *Ethics and Information Technology*. https://doi.org/10.1007/s10676-017-9425-5

Singer, P. (2009). *Animal liberation: The definitive classic of the animal movement* (Updated ed., 1st Ecco pbk. ed., 1st Harper Perennial ed). New York: Ecco Book/Harper Perennial.

Slingerland, E.G. (2014). *Trying not to try: The art and science of spontaneity* (First Edition). New York: Crown Publishers.

Sullins, J. P. (2008). Friends by design: A design philosophy for personal robotics technology. In *Philosophy and Design* (pp. 143–157). Retrieved from http://link.springer.com/chapter/10.1007/978-1-4020-6591-0_11

Turing, A. M. (2004). *The essential Turing*. Oxford University Press.

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. New York: Basic Books.

Vallor, S., & Bekey, G. A. (2017). Artificial intelligence and the ethics of self-learning robots. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 338–353). Oxford: Oxford University Press.

Wee, C. (2011). "Xin", trust, and Confucius' ethics. *Philosophy East and West*, *61*(3), 516–533.

Withers, R. (2018, April 30). I judge men based on how they talk to the Amazon Echo's Alexa. *Slate.* Retrieved July 30, 2019, from *Slate* website: https://slate.com/technology/2018/04/i-judge-men-based-on-how-they-talk-to-the-amazon-echos-alexa.html

Wong, D. (2011). *Confucian political philosophy* (Vol. 1; G. Klosko, Ed.). https://doi.org/10.1093/oxfordhb/9780199238804.003.0048

Xunzi. (2014). *Xunzi* (E. Hutton, trans.). Princeton, NJ: Princeton University Press.