The Confucian Matador: Three Defenses Against the Mechanical Bull

Tom Williams
Colorado School of Mines
Golden, CO
twilliams@mines.edu

Qin Zhu
Colorado School of Mines
Golden, CO
qzhu@mines.edu

Ruchen Wen
Colorado School of Mines
Golden, CO
rwen@mines.edu

Ewart J. de Visser
US Air Force Academy WERC
Colorado Springs, CO
ewartdevisser@gmail.com

ABSTRACT

It is critical for designers of language-capable robots to enable some degree of moral competence in those robots. This is especially critical at this point in history due to the current research climate, in which much natural language generation research focuses on language modeling techniques whose general approach may be categorized as "fabrication by imitation" (the titular mechanical "bull"), which is especially unsuitable in robotic contexts. Furthermore, it is critical for robot designers seeking to enable moral competence to consider previously under-explored moral frameworks that place greater emphasis than traditional Western frameworks on care, equality, and social justice, as the current sociopolitical climate has seen a rise of movements such as libertarian capitalism that have undermined those societal goals. In this paper we examine one alternate framework for the design of morally competent robots, Confucian ethics, and explore how designers may use this framework to enable morally sensitive human-robot communication through three distinct perspectives: (1) How should a robot reason? (2) What should a robot say? and (3) How should a robot act?

KEYWORDS

Robot Ethics, Confucian Ethics, Moral Communication

ACM Reference Format:

Tom Williams, Qin Zhu, Ruchen Wen, and Ewart J. de Visser. 2020. The Confucian Matador: Three Defenses Against the Mechanical Bull. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3319502.3378181

HRI '20, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-6746-2/20/03…$15.00
https://doi.org/10.1145/3319502.3378181

1 INTRODUCTION

Interactive robots integrated into human social environments must be able to reason, act, and communicate in a way that comports with human standards of morality; in short, they must demonstrate *moral competence* [53]. The need for moral competence is especially true for language-capable robots. To see why, it is necessary to consider current trends in the state of the art of natural language generation. Much current work on natural language generation (NLG) is broadly focused on *language modeling*. In this paradigm, NLG is not typically performed to communicate some particular fact or disposition for some particular reason, but is instead an exercise in blindly spouting whatever text is most statistically likely (or most likely to be rewarded) in the current dialogue context. This behavior is a clear example of *bullshitting* in the formal sense described by Frankfurt [28]: the "agent" (or rather, model) is generating text (1) without regards for the veracity of what is being said (2) to convince the listener of some attribute, sentiment, disposition, or state of affairs (in this case, simply that it is deserving of reward).

We call this particular form of language generation "fabrication by imitation" because language models are typically trained only for predictive accuracy with respect to some reference texts. Numerous voices both in academia and the press have expressed concern not only that language generation models in this category, like GPT-2 (Generative Pre-trained Transformer 2) [66], could be used explicitly for the purposes of a particularly pernicious form of fabrication by imitation, i.e., plagiarism [27]. Moreover, even when used for ostensibly benign purposes, GPT-2 will generate text that comes perilously close to accidental plagiarism, generating sentences and paragraphs drawn verbatim from online sources without citation [39].

Fabrication by imitation is an effective dialogue strategy in many domains of interest in current chatbot research. For example, a customer service bot seeking to answer questions to prospective diners may be quite effective simply by parroting the types of responses that human customer service agents might give, since the agent can be reasonably expected to receive a narrow range of questions regarding, e.g., hours, location, ability to accommodate dietary restrictions, etc., because answers to these questions can be easily lifted by statistical models from previous conversations, and because these answers are likely to remain highly stable, due to lack of changing context. In fact, human operators in phone-based customer service often are already operating according to a script written by someone else. Note, however, that in this sort of domain there's typically limited potential for severe negative consequences if the agent's answers are incorrect: a plausible but erroneous response will at worst result in a slightly frustrated customer upon arrival to the restaurant.

Agents operating in safety-critical environments *cannot* bullshit (again, using this term only in its formal sense). Bickmore et al. [8] demonstrate the importance of this in the context of medical advice chatbots, and highlight that in such domains, verbal responses that are inaccurate or that even accidentally imply incorrect information can literally result in patient death. The majority of domains of interest in human-robot interaction are safety critical, due to safety or privacy concerns grounded either in the environment, as in healthcare [25], space [64], and urban search and rescue [58], or in the expected user population, as in child-robot interaction [15] or eldercare robotics [29, 76].

Moreover, for language-capable robots, *every* context is potentially safety critical due to robots' unique persuasive power. Recent work has clearly demonstrated robots' ability to persuade humans [5, 10, 17, 34, 46, 56, 62, 67, 70, 77–79, 83]. While this persuasion can be leveraged for positive purposes (see, e.g., work on persuasive robotics for exercise coaching [83]), there clearly also exists the possibility for persuasion towards negative ends. Critically, this potential for negative persuasion exists even for robots designed with good intentions.

Recent work [43, 44, 88], for example, suggests that when robots simply request clarification of ambiguous but unethical commands, humans not only incorrectly infer that the robot would be willing to perform the impermissible action, but also subsequently rate that action as more permissible in their own opinion. This supports previous observations of "ripple effects" in which robots' behaviors unintentionally create new social norms between human teammates [50, 79].

If robots, through benign language patterns, can influence peoples' moral beliefs, then robots may have the ability for significant harm to their human teammates' moral ecosystem even on moral issues they know nothing about. Specifically, it may be the case that by failing to appropriately respond to immoral language (be it by failing to refuse an inappropriate command or by failing to object to racist or xenophobic statements made in its presence) the robot may be seen as condoning the immoral actions, dispositions, or sentiments at the heart of such language, and by doing so, may unintentionally and implicitly exert negative influence on their moral ecosystem. It is thus critical that robots be designed to carefully and thoughtfully generate moral language.

So how can robot designers go about effectively designing morally sensitive language-capable robots? (And do so in a way that avoids succumbing to the problematic prevailing trend in AI, i.e., the titular mechanical "bull"?) We argue that there are three distinct perspectives that can be taken to achieve this goal.

(1) Designers may use planning and reasoning algorithms and data structures that are informed by a desired moral framework.
(2) Designers may enable robots to generate moral language that explicitly espouses moral principles from a desired moral framework.
(3) Designers may enable robots to perform communicative actions that may be justifiable under a desired framework, even without operating through the first two perspectives.

The specific ways that these perspectives are employed will of course differ depending on the moral framework guiding designers. The vast majority of work on enabling autonomous moral agents and morally competent robots has been grounded in a deontological ethics (see, e.g., [12, 73]), where the morality of an action is based on whether the action is inherently right or wrong based on a set of (ostensibly) universalizable rules [1]. As pointed out by Rosemont and Ames [68], this is potentially problematic due to the emphasis in this framework (as well as in other common Western ethical frameworks) on maintenance first and foremost of *autonomous individualism* [18], in which individual liberty is prized above all else, and which thus values selfishness over attentiveness to others' needs [51]. This focus on autonomous individualism has led to a problematic movement in current Western culture (exemplified through political movements like libertarian capitalism) in which preservation of complete individual freedom is used as moral justification for curtailing social justice and equality, thus allowing for ostensible moral justification for a lack of appropriate care for others [68]. If these are values that we as roboticists wish to see in our society, then our robots must also be designed from a philosophical perspective that supports those ideals.

In this paper, we specifically examine how the three perspectives described above can be employed under a Confucian ethics framework, in particular, a Confucian role ethics framework. We choose Confucian role ethics as our exemplar moral framework for four primary reasons: (1) it is uniquely well suited to robotics due to its focus on adherence to hierarchically structured relational roles (which will necessarily govern how robots will fit into their unique socio-technical niche within human society); (2) there exist informative perspectives within recent Confucian ethics literature on social agents' duty to perform moral remonstration; (3) Confucian role ethics focuses on moral cultivation and consideration of others in determining morality, rather than on autonomous individualism; and (4) a discussion of Confucian ethics may be uniquely informative for the HRI community due to the countervailing focus in the community on Western ethical theories, especially norm-based theories such as deontology.

2 CONFUCIAN ETHICS

2.1 Key Tenets

Confucian ethics is well known for its focus on cultivating virtues in various relationships. For Confucians, what is most important for any human is to cultivate the moral self. The primary purpose of living in a Confucian life is "character-building" and self-cultivation, which serves as the "starting point and source of inspiration of character building" [80, p. 27]. Self-cultivation is necessary as both a good society and a righteous government are founded on the moral perfection of humans [16]. Self-cultivation is also possible as humans are malleable and born with the same potential to become good [57]. Hence, one's moral development depends very much on their own efforts to actively engage in learning, including learning from and interacting with others in social settings.

For Confucians, virtues are all cultivated in one's own interactive, social relationships with others. We as humans all assume different social roles (e.g., daughter, parent, professor, colleague). These social roles not only describe our relationships with others but also provide normative guidance for how to live these roles well (e.g., what a good parent might look like) [2]. Early Confucians all acknowledged that there are five cardinal role-relationships central to a Confucian society: the relationships between parents and children, husbands and wives, older and younger siblings, rulers and ministers, and friends [19]. Appropriately and continuously living these role-relationships can lead to the development of corresponding virtues (e.g., the relationship between friends can lead to the development of the virtue of faithfulness). Moreover, as particularly argued by Mencius, the moral sensitivity and concerns developed in these cardinal role-relationships can be naturally extended toward others in the society whose relationships with us might be weaker (e.g., co-nationals, strangers). Of the five cardinal role-relationships, friendship is a special one as it serves as "a bridge between the role-relations found in the family and more public roles found within broader society" [19, p. 124].

What is the most crucial for the cultivation of virtues in social relationships is self-reflection. Therefore, to be a good person, one needs to frequently reflect on their role-relationships with others, their own contribution to such relationship building, and whether these relationships are beneficial for cultivating their own virtues. The Confucian model for moral development consists of three components: observation, reflection, and practice [87]. Humans are advised to observe and reflect how themselves and others act and interact in society and integrate and test the reflective learning experience into future actions. Through the reiteration of observation, reflection, and practice, one's moral development level moves from beginner, through developing learner, and finally to *junzi* 君子 (superior person), and sage [49].

In the Confucian tradition, rituals provide an important venue for people to be aware of and sensitive to social roles and associated moral responsibilities. In ritual practices, people get the chance to reflect on their connections with the ancestors and their own roles They will also reflect on whether they have lived their assigned social roles well (e.g., whether I am a good parent) and whether their ancestors would be satisfied with their performance if they were still alive [65]. Rituals often provide a structured context in which everyone's social roles and associated normative expectations are clearly defined. By taking a ritual perspective to examine the everyday interactions in the society, Confucians may argue that it is crucial to be aware of attitudes, motivations, and values underlying routine human interactions (e.g., daily formal greetings) rather than the presentation of these interactions (e.g., simply saying please or hello). Self-reflection in the cultivation of the moral-self often requires and cultivates practical wisdom.

The *Analects* includes stories which are often piecemeal, unorganized, and sometimes quite "difficult" to reason about or understand. Western philosophers including Kant have argued that Confucian ethics cannot be considered as philosophy and Confucian classics such as the *Analects* lack systematic philosophical theories. However, philosophers such as Lai [49] argue that the Analects should be read as a manual of moral decision-making, a log, as it were, of Confucius' conversations with others [49]. Thus, readers of the Analects must extract possible reasons for why Confucius would make certain decisions in specific situations. Rather than appealing to pre-determined principles, early Confucians advocated for a moral particularism and argue that the moral actors need to exercise their imaginations to discern diverse factors and constraints present in moral situations. Readers need to "extract" from moral situations possible reasons for certain actions. Therefore, such active and reflective engagement with Confucian classics requires and cultivates moral imagination and practical wisdom.

To better understand Confucian ethics we will now briefly describe some of its recent popular interpretations, including that which we have chosen to explore in this paper, *Confucian role ethics.*

2.2 Interpretations of Confucian Ethics

Scholars of Confucian ethics have pointed out that Confucian ethics is a vision of moral life rather than an ethical theory in and of itself, and as such, it does not necessarily directly "compete" with traditional ethical theories [68]. Accordingly, to better understand Confucian ethics, scholars have sought to interpret it *through the lens* of traditional families of ethical theories, including consequentialism, deontology, care ethics, virtue ethics,and role ethics [54]. In this section we will briefly summarize some of these interpretations and how they differ both respect to each other, and from other ethical theories that also fall into those categories. Specifically, we will focus on care ethics, virtue ethics, and role ethics, due to the problems identified above with respect to autonomous individualism that come with starting from more traditional Western ethical theories such as deontology and consequentialism.

2.2.1 Confucian Ethics as Care Ethics. Care ethics is a feminist philosophy that emphasizes concern for others over protection of autonomy and adherence to norms [36]. Scholars such as Pang-White [61] have argued that although traditional Confucianism is sometimes associated with a gender-oppressive patriarchal hierarchy (cf. [37]), Confucian ethics can be interpreted as a form of care ethics due to the focus within Confucian ethics on *ren* 仁 (benevolence, goodness, humaneness). Pang-White [61] argues specifically that care ethics is better able to capture Confucian ethics than deontology or consequentialism due to the unique emphases placed by both Confucian ethics and care ethics on (1) affectionate bonds between people and the importance of moral feeling, (2) contextually sensitive application of *ren*-care over algorithmic adjudication, and (3) the inseparability of private and public spheres (e.g., the family unit and the government).

2.2.2 Confucian Ethics as Virtue Ethics. Virtue ethics is a normative ethics that emphasizes moral character and facets thereof that people may seek to cultivate [40], and that prioritizes a person's moral worth over the rightness of their actions [71]. Confucian ethics is commonly cast as a virtue ethics [84] due to its focus on cultivation of the moral self and on internalizing virtues commensurate with one's identity, e.g. through seeking to cultivate the ideal of *junzi*, which [2] translate as the "exemplary person", and the natural comparison between this and the *eudaimonia* (flourishing) espoused in Aristotelian virtue ethics.

2.2.3 Confucian Ethics as Role Ethics. In reaction to the interpretations described above, several modern philosophers have sought to instead interpret Confucian ethics through a substantially different lens, as a "role ethics" in which people are seen as "role-bearers" rather than candidate rights-holders [68, p.8]. This view combines many of the advantages of the preceding perspectives: like care ethics, role ethics emphasizes relational care over the autonomy of individuals; and like virtue ethics, role ethics emphasizes moral self-cultivation over adherence to rules. As Rosemont and Ames [68, p.9] argue, this is an advantageous exchange in both cases, as libertarianism capitalism in the United States has demonstrated that solely emphasizing individual autonomy and norm adherence comes at a cost of reduced equality and social justice.

Instead, role ethics emphasizes the roles that agents play *with respect to one another*, both descriptively (capturing the nature of the different relational roles that humans objectively play with respect to one another) and proscriptively (capturing the normative responsibilities that dictate appropriate conduct with respect to those roles) [68, p.12]. Confucian role ethics thus not only provides a new perspective for interpreting the moral precepts laid out in *Analects* but also provides a valuable new ethical framework for understanding and reacting to the modern world (e.g. with respect to the rise of libertarian capitalism).

In this paper, we describe the insights that Confucian role ethics may provide to robot designers through multiple perspectives at different levels of analysis. To do so, we will begin by defining these disparate perspectives, from a starting point grounded in current frameworks for enabling morally competent robots.

3 CONFUCIAN PERSPECTIVES ON MORALLY COMPETENT ROBOTS

Malle and Scheutz [53] delineate a set of four key requirements for enabling morally sensitive robots, presented here verbatim:

(1) A system of norms and the language and concepts to communicate about these norms;
(2) Moral cognition and affect;
(3) Moral decision making and action; and
(4) Moral communication.

In contrast, a Confucian perspective towards enabling moral competence might require alternative requirements:

(1) Whereas the traditional requirements described above require only a system of norms and the language and concepts to communicate about those norms, a set of Confucian requirements might additionally or alternatively require a system for representing the relationships between members of the robot's social context (including itself); a way to represent the (possibly context-sensitive) roles the robot plays in those relationships; a way of either specifying what actions are viewed as benevolent with respect to each of those roles or a way of associating moral norms with those roles; and a set of language and concepts to communicate about those roles and relationships.
(2-3) Whereas the traditional requirements described above require a way to use the required set of norms to judge and respond emotionally to norm violations and to make morally-sensitive decisions, a set of Confucian requirements might naturally require these judgments, emotional responses, and decisions to also be sensitive to the system of roles and relationships described above.
(4) Moral Communication (ability to explain, justify,criticize, etc. norm violations) – ability to perform these actions using Confucian rationale (e.g., grounded in roles)

These requirements can be satisfied from a Confucian perspective in three ways:

(1) All four of these requirements can be satisfied in accordance with a Confucian perspective by using data structures and algorithms directly informed by Confucian ethical principles.
(2) The last of these requirements can be satisfied in accordance with a Confucian perspective by having the robot directly communicate or ground its communications in Confucian ethical principles; regardless of whether or not the decision to communicate that information is made using CE-theoretic data structures and algorithms.
(3) More holistically, the robot's behavior, and the behavior it encourages in others, can be assessed as to whether that behavior or encouraged behavior conforms with Confucian ethical principles; again, regardless of whether the behavior or means of encouraging others behavior is chosen according to Confucian ethical principles or whether the robot chooses to explicitly communicate Confucian ethical principles.

3.1 The First Way: How Should a Robot Reason?

The first Confucian perspective designers may take when designing morally sensitive robots is to design data structures and algorithms that are grounded in Confucian concepts. For example, designers may seek to achieve role-based equivalents to traditional norm-based requirements for morally competent robots. Specifically, Malle and Scheutz [53]'s first requirement requires, in part, a system of moral norms that may subsequently be used for making moral judgments and moral decisions, and which may be used as the basis for moral communication.

In contrast, robot designers operating within a Confucian perspective may additionally or alternatively require a system for representing the relationships between members of the robot's social context (including itself); a way to represent the (possibly context-sensitive) roles the robot plays in those relationships; and a way of either specifying what actions are viewed as benevolent with respect to each of those roles or a way of associating moral norms with those roles. These Confucian-inspired data structures and knowledge representations could then in turn allow for satisfaction of Malle and Scheutz [53]'s other requirements within a Confucian perspective, i.e., by allowing for the development of algorithms for moral judgment, moral decision making, and moral communication that are grounded in these role-based representations.

Robot designers seeking to explore this perspective should begin by determining what roles and relationships they will need to represent within their application domain. While these roles and relationships may substantially differ between application domains, we believe that all social robots will need to represent a "relational core": a variation on the five cardinal relationships of Confucianism, reoriented towards social robots.

Within Confucianism, five key relationships are espoused: ruler-minister, father-son, husband-wife, older-younger, and friend-friend. Of these, only friend-friend is clearly applicable to robots; as such, we propose the following alternative set of cardinal relationships for human-robot interaction: supervisor-subordinate, owner-ownee, teammate-teammate, adept-novice, and friend-friend, as shown in Fig. 1 and originally discussed in [89][1]. These relationships may serve as a relational core for human-robot interaction that may be extended by designers as required for their target domain. Once an appropriate set of relationships for a given application domain have been determined, robot designers must make a series of supplemental decisions (or provide mechanisms to allow robots to make or learn to make these decisions for themselves).

First, it must be determined in what contexts and with respect to whom does the robot hold these relational roles. Depending on application domain and the type of relational role, it may be more appropriate to encode this knowledge with respect to specific individuals (the relational role of *owner*, for example, will typically hold with respect to a specific individual or set of individuals), while in others it may be more appropriate to encode this knowledge with respect to a class of identifiable individuals (a service robot in a public context, for example, may need to hold anyone identifiable as a customer in the relational role of *supervisor*).

Next, it must be determined what actions are considered benevolent for a given relational role, or how benevolence can be dynamically assessed. In making this decision, designers may incorporate other moral frameworks in order to create a hybrid moral reasoning system. This benevolence may, for example, be assessed using some form of utilitarian calculus, or through a set of deontic norms (it is less clear how virtue ethics or care ethics could be appropriately computationalized, although cf. Kuipers' suggestion towards case-based models of virtue ethics [48]). Critically, however, even if another moral framework is incorporated at this point in order to perform this assessment of benevolence, the use of relational roles as the bedrock representation will help to ensure a Confucian basis

CARDINAL RELATIONSHIPS FOR HUMAN-ROBOT INTERACTION

owner ⟷ **ownee**
An asymmetrical relationship on the basis of ownership, as advocated by Bryson [14]. Example: an elderly person and their personal service robot [69]

supervisor ⟷ **subordinate**
An asymmetrical relationship within an organizational structure; the predominant asymmetrical relationship in task-based HRI [38, 74, 85]. Example: an employee tasking a robot to perform a custodial task [11].

adept ⟷ **novice**
An asymmetrical relationship on the basis of competence or expertise [24]. Example: a child working with an intelligent robot tutor [47].

teammate ⟷ **teammate**
A symmetrical relationship within an organizational structure; the predominant symmetrical relationship in task-based HRI [74, 85]. Example: A Robonaut operating alongside an astronaut in a maintenance task [26].

friend ⟷ **friend**
A symmetrical social relationship (note arguments both for [20] and against [21, 22, 72] the development of such relationships in practice). Example: a hospitalized child playing with a robot companion [6].

Figure 1

for moral decision making and provide an opportunity for straightforward generation of moral explanations grounded in Confucian principles, as discussed in Section 3.2.

Case Study

To explore how design through this perspective might play out, we present some preliminary representations that we imagine could be designed through a role-based perspective, building on recent knowledge representation work outside the context of moral reasoning [82]. A role-based approach to moral competence must be sensitive to dynamic change. A role-based approach requires robots to have prior knowledge of roles assumed in specific contexts, and the relationship the robot has with their human interactants. In real-life scenarios, contexts change dynamically, and the roles and relationships of the agents thus also change dynamically. In addition, a role-based approach to moral competence must be sensitive to uncertainty. A robot cannot be sure that it has the most accurate knowledge of its current context at all times.

Given these two requirements, one candidate representation for role-sensitive norms might be to extend previously presented norms representations grounded in Uncertain First-Order Logic [82], which allows for reasoning under uncertainty over first-order logic formulas [59, 60].

[1]It still may be valuable for robots to *recognize* the traditional Confucian cardinal relationships, even if they are unable to participate in them. However, these traditional roles may need to be modified in terms of the set of states and actions viewed as good or benevolent with respect to those roles. For example, it may be valuable to recognize the relationship of spouse-spouse, but (as indicated through this rephrasing) the associations that accompany that relationship would need to be reevaluated in order to avoid the patriarchal gender norms often associated with classical Confucianism.

Within this framework, an actor r might be represented as:

$$r = Role(Ag)$$

where *Role* denotes the role that the agent is assumed in a specific context (e.g., teacher, doctor), and *Ag* denotes an agent. For an agent whose name is Jake and having a role as a student, this actor can be represented as "$Student(Jake)$".

A *Norm* $\mathcal{N}$ can then be defiend as an expression of the form:

$$\mathcal{N} := \mathcal{R} \wedge C \Rightarrow \mathcal{A}$$

where a relationship $\mathcal{R} = \{r_1, \ldots, r_n\}$ is a non-empty set of actors, C represents a possibly empty set of contextual conditions and $\mathcal{A}$ represents a role-suitable action that can be inferred from a relationship $\mathcal{R}$ and contextual conditions C. A norm "A police officer X who is on-duty can issue tickets to a driver Y who is speeding" can be represented as" $\{Police(X), Driver(Y)\} \wedge \{on\_duty(X) \wedge speeding(Y)\} \Rightarrow issue\_tickets(X, Y)$".

An Uncertain First-Order *Belief-Theoretic Norm* $\mathcal{N}_{[\alpha_\mathcal{N}, \beta_\mathcal{N}]}$ can then be defined as an expression of the form:

$$\mathcal{N}_{[\alpha_\mathcal{N}, \beta_\mathcal{N}]} := \mathcal{R}_{[\alpha_\mathcal{R}, \beta_\mathcal{R}]} \wedge C_{[\alpha_C, \beta_C]} \Rightarrow \mathcal{A}_{[\alpha_\mathcal{A}, \beta_\mathcal{A}]}$$

where each norm and proposition are associated with their own Dempster-Shafer theoretic uncertainty intervals [75], with lower probability bound α and upper probability bound β.

One limitation of this approach (which follows quite closely off of previously presented work in norm representation) is that the roles represented here are not inherently *relational* as we might desire from a Confucian ethics approach. Accordingly, to represent context-sensitive relationships and roles, we might modify the above representations to represent relational roles as bidirectional relational pairs: $\{role\_of(A, B), inverse\_role\_of(B, A)\}$, allowing for relational inference rules such as the following to be formulated:

$$\{teach\_course(X, Z) \wedge enroll\_in\_course(Y, Z)\}$$
$$\Rightarrow \{teacher\_of(X, Y), student\_of(Y, X)\}$$

Inferring a relational role in this way based on context would then, when used in conjunction with the previously specified belief theoretic norms, allow an agent to determine what actions are benevolent according to its current role, even in dynamic contexts and unde conditions of uncertainty and ignorance.

3.2 The Second Way: What Should a Robot Say?

The second Confucian perspective designers may take when designing morally sensitive robots is to design behaviors that explicitly reference Confucian principles. There are many contexts in which language-capable robots may need to engage in moral communication with regards to some moral decision or position. Aschenbrenner [3] lays out five main categories of such moral judgment, which we here expand on and connect to language capable robots:

(1) **Remonstrance:** in which a robot must explain (a) its reason for rejecting a command explicitly or implicitly delivered by an interlocutor (refusal of strong uptake) or (b) its disagreement with a proposition explicitly or implicitly asserted by an interlocutor (refusal of weak uptake).
(2) **Accusation:** in which a robot levels moral criticism against a party deemed to be acting immorally, either through observed performance, proposal, requesting, or condoning of such action.
(3) **Correction:** in which a robot not only levels an accusation but furthermore issues moral sanction against the accused.
(4) **Intercession:** in which a robot argues on behalf of an accused violator.
(5) **Inculpation:** in which a robot apologizes for or otherwise acknowledges its own immoral action.
(6) **Exculpation:** in which a robot forgives or absolves an accused violator.

When performing any of these actions, robots have the opportunity to ground their justifications in different moral theories when explaining *why* they are remonstrating, accusing, correcting, interceding on behalf of, inculpating, or exculpating a particular agent.

Consider, for example, a robot asked to retrieve a wallet that has been left unattended on a nearby desk. A robot rejecting this command might deferentially highlight different moral concepts in its moral remonstration (cf. [81]). For example, the robot could directly highlight the would-be-violated norm at the heart of the justification by saying "I cannot do that because that would be stealing, and stealing is wrong." This type of phrasing can be viewed as being grounded in Kantian categorical imperatives: absolute moral norms that the agent (and everyone else) is supposed to comply with across all moral situations regardless of the consequences due to universal principles of *right action*.

Alternatively, the robot could instead highlight relational roles that might underlie its justification by saying "I cannot do that because that wallet belongs to my friend Sean, and a good friend would not do that." This type of phrasing can be viewed as grounded in a Confucian role ethics due to its emphasis of the friend-friend relationship (cf. [86]).

An approach following this perspective would be facilitated by the first perspective: the existence of knowledge representations for encoding relational roles provides a natural opportunity for using those representations when generating moral language. However, these first two perspectives are in fact independent of one another. A robot with role-oriented knowledge representations could still in principle generate norm-based moral communications; and a robot lacking role-oriented knowledge representations could still in principle generate role-based moral communications. This approach could be pursued if it were empirically determined that the representations and algorithms that enabled most effective or efficient moral reasoning was not the most effective for generating effective moral communications. If this were the case, designers would need to make a choice between effectiveness (of either moral reasoning or moral communication) and transparency (enabling robots' true moral reasoning procedures to be accurately reflected in its moral communications). We also stress here that while for the purposes of this paper we are turning our attention specifically towards role ethics in this paper, for this perspective care- or virtue-ethics approaches may also have benefit.

Use of different perspectives for moral reasoning and moral communication, however, may come at a cost. *Transparent* communication of moral principles (beyond merely following those principles) can be particularly important. Research on human moral and social norms suggests that in order for norms to remain norms, they must be clearly communicated between community members [7, 30, 31]. Thus, even for robots espousing role-based tenets, compliance with and attention to those tenets may itself be a norm that robots may need to clearly communicate in order for it to take hold in human teammates.

Case Study

As a concrete example of this perspective, we consider the work of Wen et al. [81]. In that work, researchers conducted a preliminary comparison of moral remonstrations grounded in roles versus norms for a variety of utterance types (i.e., questions, refusals, and rebukes), in the context of proposed violations of norms of honesty (i.e, cheating). For example, for refusals, the researchers contrast "That would be cheating and cheating is wrong. I won't tell you." (a norm-based refusal grounded in the inherent wrongness of the proposed action) vs. "A good instructor wouldn't do that. I won't tell you." (a role-based refusal grounded in the disconnect between the proposed action and the practices that manifest *ren* for the role the robot holds with respect to the agent with whom it is remonstrating). Through this experiment, the researchers produced preliminary results suggesting that role-based language may be better at encouraging perception of an agent as successful within its social role, but may be less effective at encouraging other desiderata such as mindfulness (as norm-based language language may provoke more extreme and immediate emotional responses).

3.3 The Third Way: How Should a Robot Act?

The third Confucian perspective designers may take when designing morally sensitive robots is to analyze whether a robot's behavior exhibits core Confucian moral principles, regardless of framing of Confucian ethics with respect to other ethical frameworks, and moreover regardless of whether its underlying moral reasoning framework is designed to represent Confucian principles or whether the robot is designed to espouse Confucian principles in its communication. As an example, let us continue within the context of moral communication. How might robot designers assess candidate approaches towards moral communication through a Confucian perspective, without focusing on Confucian representations or Confucian linguistic behaviors?

As discussed in Section 2.1, Confucian ethics emphasizes the ability for moral cultivation, and the benefits of self-reflection in encouraging such cultivation. Accordingly, robot designers may ask "Does this robot design encourage the moral cultivation of its users, e.g., through self-reflection and cultivation of the heart of shame?" Critically, the answer to this question may very well be "yes" even without adherence to the first two perspectives we have presented. That is, a robot may well be able to achieve goals espoused by Confucian ethics without knowledge representations explicitly designed to reflect Confucian principles, and without explicitly including such principles in its moral communications.

Case Studies

One initial foray into Confucian robot ethics has been made by Philosopher JeeLoo Liu, who presents a Confucian analogue to Asimov's three laws of robotics [52].

CR1 A robot must first and foremost fulfill its assigned role.
CR2 A robot should not act in ways that would afflict the highest displeasure or the lowest preference onto other human beings, when other options are available.
CR3 A robot must render assistance to other human beings in their pursuit of moral improvement, unless doing so would violate [CR1] and [CR2]. A robot must also refuse assistance to other human beings when their projects would bring out their evil qualities or produce immorality.

Nearly 70 years of research (along with Asimov's stories themselves [4]) have demonstrated that seeking to enable morally competent robots by programming them to comply with a set of limited set of universal principles such as Asimov's Three Laws is unlikely to be successful due to both their computational intractability and their ambiguity (which leads to the unexpected outcomes that make Asimov's stories so entertaining in the first place) [55]. This suggests that rules such as those proposed by Liu are unlikely to be useful *if employed from a First Way perspective*. In contrast, sets of principles such as these may indeed prove useful from a *Third Way* perspective (cf. [55]); they may serve as useful Confucian-oriented design guidelines for robot designers. From this perspective designers may ask, *regardless of the robot's programming*: Does the robot behave in a way that fulfills its assigned role? Does the robot behave in a way that avoids inflicting the highest displeasure or lowest preference onto other human beings? And does the robot behave in a way that encourages the moral cultivation of others?

Furthermore, the Case Study presented as an example of the Second Way may also be used to illustrate the perspective of the Third Way. Specifically, Wen et al. [81]'s consideration of remonstrative phrasings that differ according to illocutionary point (i.e., utterance types) allows them to explore how phrasing might be used to accomplish different Confucian goals even without explicitly incorporating Confucian principles (i.e., roles) into said phrasing. While Wen et al. [81] varied phrasings in this way in order to ensure that their results were not limited in scope to one particular utterance phrasing, future work could investigate how intentional variation of that phrasing (cf. [42, 45]), regardless of whether roles or norms are highlighted in remonstrations, could be used to achieve moral goals. For example, remonstrations phrased as questions rather than rebukes could be more effective at encouraging self-reflection and moral self-cultivation, even if they are no more effective at effecting immediate behavioral changes.

4 CLOSING THOUGHTS

In this paper, we have demonstrated how three distinct perspectives (*How should a robot reason? What should a robot say? How should a robot act?*) may be used in the design of morally competent language-capable robots, through a process that is careful and intentional, in contrast to the "fabrication by imitation" approach that is in vogue today. Critically, all three of these perspectives serve as a foil vis-a-vis fabrication by imitation: a designer taking the first perspective is guaranteeing that communication is grounded in some verifiable reasoning; a designer taking the second perspective is exerting direct control over the robot's messaging in order to ensure that the robot's utterances not only do not violate moral principles but rather that they directly espouse them; and a designer taking the third perspective is ensuring that the standards used to measure the quality of the designed robot system are grounded in morally justifiable values rather than solely in imitative capacity. Moreover, we have demonstrated how these perspectives may be used to enable robots that comport with the principles of Confucian role ethics, which may be uniquely suited to encourage the social goals of care, equality, and social justice that have been degraded within today's sociopolitical climate.

A number of lingering concerns remain. Even if robot designers follow the three principles discussed in this paper, they must be mindful of how people might accept or trust the robot's reasoning, communication and behavior. Three issues come to mind in this regard. The first is that people might not perceive the robot to have agency or experience feelings [9], two dimensions that are critical to mind perception [32], which is essential for perceived moral agency [33]. The second is the notion of the face threatening act (FTA) proposed as part of human politeness strategies [13]. Imposing moral behavior from a robot might heighten the FTA because it threatens the autonomy, known as the negative face, of interaction partners [35]. The expression of moral judgments and actions may thus require sophisticated politeness delivery strategies to dampen a heightened FTA [41, 63], unless designers seek to explicitly challenge users' valuation of autonomy. Lastly, if robots enter in longer-term relationships with humans, the notion of relationship equity may become important [23]. Relationship equity is the goodwill that exists in a relationship between two actors as a result of the difference between relationship costs and benefits. A moral correction from a robot could either be a cost or a benefit to a human partner depending on how justified this correction is perceived. For example, if the human partner shares the underlying value behind the moral correction it might be perceived as a benefit. If the human partner perceives that correction as, for example, coming from the robot's designer or distributor or disagreeing with the correction, it might be considered a cost. Future research might examine how perceptions of moral reasoning, communication and behavior impact longitudinal human-robot interaction.

ACKNOWLEDGMENTS

The authors would like to thank Elizabeth Phillips for her helpful comments. This work was funded in part by National Science Foundation grant IIS-1909847.

REFERENCES

[1] Larry Alexander and Michael Moore. 2007. Deontological ethics. (2007).

[2] Roger T Ames and Henry Rosemont. 2010. *The analects of Confucius: A philosophical translation.* Ballantine books.

[3] Karl Aschenbrenner. 1971. Moral Judgment. In *The Concepts of Value.* Springer.

[4] Isaac Asimov. 1942. Runaround. *Astounding Science Fiction* 29, 1 (1942), 94–103.

[5] Ilaria Baroni, Marco Nalin, Mattia Coti Zelati, Elettra Oleari, and Alberto Sanna. 2014. Designing motivational robot: how robots might motivate children to eat fruits and vegetables. In *Int'l Symp. Robot and Human Interactive Communication.*

[6] Tony Belpaeme, Paul Baxter, Robin Read, Rachel Wood, Heriberto Cuayáhuitl, Bernd Kiefer, Stefania Racioppa, Ivana Kruijff-Korbayová, Georgios Athanasopoulos, Valentin Enescu, et al. 2013. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction* 1, 2 (2013), 33–53.

[7] Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

[8] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Jour. Medical Internet Research* (2018).

[9] Yochanan E Bigman, Adam Waytz, Ron Alterovitz, and Kurt Gray. 2019. Holding Robots Responsible: The Elements of Machine Morality. *Trends in cognitive sciences* 23, 5 (2019), 365–368.

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

[11] Gordon Michael Briggs and Matthias Scheutz. 2013. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Twenty-Seventh AAAI Conference on Artificial Intelligence.*

[12] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems* 21, 4 (2006), 38–44.

[13] Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage.* Vol. 4. Cambridge university press.

[14] Joanna J Bryson. 2010. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (2010), 63–74.

[15] Ginevra Castellano and Christopher Peters. 2010. Socially perceptive robots: Challenges and concerns. *Interaction Studies* 11, 2 (2010), 201.

[16] Chung-ying Cheng. 2004. A theory of Confucian selfhood: Self-cultivation and free will in Confucian philosophy. In *Confucian ethics: A comparative study of self, autonomy, and community*, K. Shun and D. Wong (Eds.).

[17] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *International conference on Human-Robot Interaction (HRI).* ACM.

[18] John Christman. 2008. Autonomy in moral and political philosophy. *Stanford encyclopedia of philosophy* (2008).

[19] C. Cottine. 2020. That's what friends are for: A Confucian perspective on the moral significance of friendship. In *Perspectives in role ethics: Virtues, reasons, and obligation*, T. Dare and C. Swanton (Eds.). 123–142.

[20] John Danaher. 2019. The philosophical case for robot friendship. *Journal of Posthuman Studies* 3, 1 (2019), 5–24.

[21] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L Walters, Kheng Lee Koay, and Iain Werry. 2005. What is a robot companion-friend, assistant or butler?. In *2005 IEEE/RSJ international conference on intelligent robots and systems.* IEEE, 1192–1197.

[22] Maartje MA de Graaf. 2016. An ethical evaluation of human–robot relationships. *International journal of social robotics* 8, 4 (2016), 589–598.

[23] Ewart J de Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2019. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *Int'l Jour. Social Robotics* (2019).

[24] Autumn Edwards, Chad Edwards, Patric R Spence, Christina Harris, and Andrew Gambino. 2016. Robots in the classroom: Differences in students' perceptions of credibility and learning between "teacher as robot" and "robot as teacher". *Computers in Human Behavior* 65 (2016), 627–634.

[25] Baowei Fei, Wan Sing Ng, Sunita Chauhan, and Chee Keong Kwoh. 2001. The safety issues of medical robotics. *Reliability Engineering & System Safety* (2001).

[26] Terrence Fong, Illah Nourbakhsh, Clayton Kunz, Lorenzo Fluckiger, John Schreiner, Robert Ambrose, Robert Burridge, Reid Simmons, Laura Hiatt, Alan Schultz, et al. 2005. The peer-to-peer human-robot interaction project. In *Space 2005.* 6750.

[27] Errol Francke and Bennett Alexander. 2019. The Potential Influence of Artificial Intelligence on Plagiarism: A Higher Education Perspective. In *European Conference on the Impact of Artificial Intelligence and Robotics.*

[28] Harry G Frankfurt. 1986. *On bullshit.* Princeton University Press Princeton, NJ.

[29] Susanne Frennert and Britt Östlund. 2014. Seven matters of concern of social robots and older people. *International Journal of Social Robotics* 6, 2 (2014).

[30] Francesca Gino. 2015. Understanding ordinary unethical behavior: Why people who value morality act immorally. *Current opinion in behavioral sciences* 3 (2015), 107–111.

[31] Susanne Göckeritz, Marco FH Schmidt, and Michael Tomasello. 2014. Young children's creation and transmission of social norms. *Cog. Dev.* (2014).

[32] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of mind perception. *Science* (2007), 619–619.

[33] Kurt Gray, Liane Young, and Adam Waytz. 2012. Mind perception is the essence of morality. *Psychological inquiry* 23, 2 (2012), 101–124.

[34] Jaap Ham, René Bokhorst, Raymond Cuijpers, David van der Pol, and John-John Cabibihan. 2011. Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power.

In *International conference on social robotics*. Springer, 71–83.
[35] Caroline Hayes and Christopher Allan Miller. 2010. *Human-computer etiquette*. Auerbach.
[36] Virginia Held et al. 2006. *The ethics of care: Personal, political, and global*. Oxford University Press on Demand.
[37] Ranjoo Seodu Herr. 2003. Is Confucianism compatible with care ethics? A critique. *Philosophy east and west* (2003), 471–489.
[38] Pamela J Hinds, Teresa L Roberts, and Hank Jones. 2004. Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction* 19, 1 (2004), 151–181.
[39] Andrew Hoang, Antoine Bosselut, Asli Celikyilmaz, and Yejin Choi. 2019. Efficient Adaptation of Pretrained Transformers for Abstractive Summarization. *arXiv preprint arXiv:1906.00138* (2019).
[40] Rosalind Hursthouse and Glen Pettigrove. 2018. Virtue Ethics. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.).
[41] Ohad Inbar and Joachim Meyer. 2019. Politeness Counts: Perceptions of Peacekeeping Robots. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019).
[42] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. 2019. Tact in noncompliance: The need for pragmatically apt responses to unethical commands. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.
[43] Ryan Blake Jackson and Tom Williams. 2018. Robot: Asker of questions and changer of norms? *Proceedings of ICRES* (2018).
[44] Ryan Blake Jackson and Tom Williams. 2019. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In *Companion Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*.
[45] Ryan Blake Jackson, Tom Williams, and Nicole M. Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.
[46] James Kennedy, Paul Baxter, and Tony Belpaeme. 2014. Children comply with a robot's indirect requests. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 198–199.
[47] James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme. 2016. Social robot tutoring for child second language learning. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 231–238.
[48] Benjamin Kuipers. 2018. How can we trust a robot? *Commun. ACM* 61, 3 (2018), 86–95.
[49] Karyn Lai. 2007. Understanding Confucian ethics: Reflections on moral development. *Australian Journal of Professional and Applied Ethics* 9, 2 (2007), 21–27.
[50] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, and Paul Rybski. 2012. Ripple effects of an embedded social agent: a field study of a social robot in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
[51] Emmanuel Levinas. 1979. *Totality and infinity: An essay on exteriority*. Vol. 1. Springer Science & Business Media.
[52] JeeLoo Liu. 2017. Confucian Robotic Ethics. In *International Conference on the Relevance of the Classics under the Conditions of Modernity: Humanity and Science*.
[53] Bertram F Malle and Matthias Scheutz. 2014. Moral competence in social robots. In *International Symposium on Ethics in Engineering, Science, and Technology*.
[54] Sarah Mattice. 2019. Confucian Role Ethics: Issues of Naming, Translation, and Interpretation. *The Bloomsbury Research Handbook of Early Chinese Ethics and Political Philosophy* (2019), 25.
[55] Lee McCauley. 2007. The frankenstein complex and Asimov's three laws. *University of Memphis* (2007).
[56] Cees Midden and Jaap Ham. 2012. The illusion of agency: the influence of the agency of an artificial agent on its persuasive power. In *International Conference on Persuasive Technology*. Springer, 90–99.
[57] Donald J Munro. 1971. The concept of man in early China. (1971).
[58] Robin R Murphy. 2004. Trial by fire [rescue robots]. *IEEE Robotics & Automation Magazine* 11, 3 (2004), 50–61.
[59] Rafael C Núñez, Manohar N Murthi, Kamal Premaratne, Matthias Scheutz, and Otávio Bueno. 2018. Uncertain Logic Processing: logic-based inference and reasoning using Dempster–Shafer models. *Int'l Jour. Approx. Reasoning* (2018).
[60] Rafael C Núnez, Matthias Scheutz, Kamal Premaratne, and Manohar N Murthi. 2013. Modeling uncertainty in first-order logic: a Dempster-Shafer theoretic approach. In *Int'l Symp. on Imprecise Probability: Theories and Applications*.
[61] Ann A Pang-White. 2009. Reconstructing modern ethics: Confucian care ethics. (2009).
[62] Raul Benites Paradeda, Maria José Ferreira, João Dias, and Ana Paiva. 2017. How Robots Persuasion based on Personality Traits May Affect Human Decisions. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 251–252.
[63] Raja Parasuraman and Christopher A Miller. 2004. Trust and etiquette in high-criticality automated systems. *Commun. ACM* 47, 4 (2004), 51–55.
[64] Liam Pedersen, David Kortenkamp, David Wettergreen, I Nourbakhsh, and David Korsmeyer. 2003. A survey of space robotics. (2003).
[65] Michael Puett and Christine Gross-Loh. 2016. *The path: What Chinese philosophers can teach us about the good life*. Simon and Schuster.
[66] Alex Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better Language Models and Their Implications. https://openai.com/blog/better-language-models/. (Accessed on 12/10/2019).
[67] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. 101–108.
[68] Henry Rosemont and Roger T Ames. 2016. *Confucian role ethics: A moral vision for the 21st century?* Vol. 5. Vandenhoeck & Ruprecht.
[69] Nicholas Roy, Gregory Baltus, Dieter Fox, Francine Gemperle, Jennifer Goetz, Tad Hirsch, Dimitris Margaritis, Michael Montemerlo, Joelle Pineau, Jamie Schulte, et al. 2000. Towards personal service robots for the elderly. In *Workshop on Interactive Robots and Entertainment (WIRE 2000)*, Vol. 25. 184.
[70] Eduardo Benítez Sandoval, Jürgen Brandstetter, and Christoph Bartneck. 2016. Can a robot bribe a human?: The measurement of the negative side of reciprocity in human robot interaction. In *Int'l Conf. on Human Robot Interaction (HRI)*.
[71] John Santiago. 2008. Confucian Ethics in the Analects as Virtue Ethics. (2008).
[72] Matthias Scheutz. 2011. The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. *Robot ethics: The ethical and social implications of robotics* (2011), 205.
[73] Matthias Scheutz, Bertram Malle, and Gordon Briggs. 2015. Towards morally sensitive action selection for autonomous social robots. In *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.
[74] Jean Scholtz. 2003. Theory and evaluation of human robot interactions. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE, 10–pp.
[75] Glenn Shafer. 1976. *A mathematical theory of evidence*. Vol. 42. Princeton university press.
[76] Amanda Sharkey and Noel Sharkey. 2012. Granny and the robots: ethical issues in robot care for the elderly. *Ethics and information technology* 14, 1 (2012), 27–40.
[77] Michael Steven Siegel. 2008. *Persuasive robotics: how robots change our minds*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[78] Megan Strait, Cody Canning, and Matthias Scheutz. 2014. Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI)*.
[79] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Int'l Conf. on Human-Robot Interaction (HRI)*.
[80] Tu Wei-Ming. 1999. Self-cultivation as education embodying humanity. In *The proceedings of the twentieth world congress of philosophy*, Vol. 3. 27–39.
[81] Ruchen Wen, Ryan Blake Jackson, Tom Williams, and Qin Zhu. 2019. Towards a role ethics approach to command rejection. In *HRI Workshop on the Dark Side of Human-Robot Interaction*.
[82] Ruchen Wen, Mohammed Aun Siddiqui, and Tom Williams. 2020. Dempster-Shafer Theoretic Learning of Indirect Speech Act Comprehension Norms. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.
[83] Katie Winkle, Séverin Lemaignan, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2019. Effective persuasion strategies for socially assistive robots. In *International Conference on Human-Robot Interaction (HRI)*.
[84] David Wong. 2018. Chinese Ethics. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.).
[85] Holly A Yanco and Jill Drury. 2004. Classifying human-robot interaction: an updated taxonomy. In *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 3. IEEE, 2841–2846.
[86] June Ock Yum. 1988. The impact of Confucianism on interpersonal relationships and communication patterns in East Asia. *Communications Monographs* (1988).
[87] Qin Zhu. 2018. Engineering ethics education, ethical leadership, and Confucian ethics. *International Journal of Ethics Education* 3, 2 (2018), 169–179.
[88] Qin Zhu, Tom Williams, and Ryan Blake Jackson. 2018. Blame-laden moral rebukes and the morally competent robot: A Confucian ethical perspective. In *Proceedings of the Workshop on Brain-Based and Artificial Intelligence*.
[89] Qin Zhu, Tom Williams, and Ruchen Wen. 2019. Confucian Robot Ethics. *Computer Ethics-Philosophical Enquiry (CEPE) Proceedings* 2019, 1 (2019), 12.