Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-based Relational Norms

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

Zhao Han
Colorado School of Mines
Golden, CO, USA
zhaohan@mines.edu

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

***Abstract*—Language-capable robots require moral competence, including representations and algorithms for moral reasoning and moral communication. We argue for an ethical pluralist approach to moral competence that leverages and combines disparate ethical frameworks, and specifically argue for an approach to moral competence that is grounded not only in Deontological norms (as is typical in the HRI literature) but also in Confucian relational roles. To this end, we introduce the first computational approach that centers relational roles in moral reasoning and communication, and demonstrate the ability of this approach to generate both context-oriented *and* role-oriented explanations for robots' rejections of norm-violating commands, which we justify through our pluralist lens. Moreover, we provide the first investigation of how computationally generated role-based explanations are perceived by humans, and empirically demonstrate (N=120) that the effectiveness (in terms of of trust, understanding confidence, and perceived intelligence) of explanations grounded in different moral frameworks is dependent on nuanced mental modeling of human interlocutors.**

***Index Terms*—Robot Ethics, Confucian Ethics, Moral Communication**

Fig. 1. **Experimental Contexts.** We compared four norm violation responses across four contexts, in each of which the robot played fundamentally different roles, and bore fundamentally different role-obligations. The images shown above depict frames from videos shown to participants during this experiment. Each image depicts an *Action* explanation used in a different context.

I. Introduction

For language-capable robots to be successfully deployed, they require *moral competence* [1] (i.e., capabilities of reasoning, acting and communicating in accordance with a moral system) to avoid negatively impacting human moral ecosystems [2]. This is critical not only in contexts where robots pose risks of physical harm, like factory contexts and space exploration, but also in contexts where robots pose risks of emotional harm, like eldercare and childcare. Moreover, without careful design, robots stand to negatively impact the beliefs, desires, and intentions of interactants in morally consequential ways. A number of HRI studies have shown that robots have significant persuasive power, and that interactants regularly comply with robots' commands and requests [3], [4], [5]. Moreover, recent work has shown that robots can exert moral influence over the systems of moral norms that govern interactants' behavior [6], [7], [8].

Malle and Scheutz suggest that four key criteria are required for moral competence: (1) a system of moral norms (2) norm-driven moral cognition to generate emotional responses to norm violations and make moral judgements, (3) norm-driven moral decision making and action, and (4) norm-driven moral communication to generate morally sensitive language for explaining one's actions and regulating others' behaviors [1], [9]. Due to this focus on norms, recent approaches to achieving robotic moral competence [1] have predominantly relied on norm-driven Western ethical theories such as deontology, which center adherence to universalizable moral rules.

HRI researchers have argued, however, that our community needs to go beyond these ethical theories, and embrace a wider diversity of moral philosophies from disparate global cultures [10]. This is important (1) so that robots can intelligently operate within different cultures within an increasingly interconnected, globalized world [11], [10]; (2) so that robot designers can center cultures whose perspectives have been historically excluded from robot interaction design (e.g., as part of decolonial [12] or anti-racist [13] computing projects). Moreover, through this lens of *ethical pluralism* [11], it is important not only to consider different cultures' ethical frames separately, but also to create robots that simultane-

This work was supported in part by NSF grants IIS-1909847 and IIS-1849348.

ously leverage multiple ethical theories as part of their moral reasoning processes.

Recent HRI research seeking to embody an ethical pluralist approach has explored what robotic moral competence might look like through the lens of an Eastern ethical tradition — Confucian Role Ethics [14], [15], [2], [10], [16], [17], which argues that moral norms are derived from the social roles humans assume and the relationships humans have with others [18]. Two of the key elements of Confucian Role Ethics are (1) a focus on roles and relationships rather than norms (although those roles certainly come with normative expectations), and (2) a focus on the cultivation of the moral self in concert with others, including the responsibility to help others grow virtues in social interactions (rather than merely avoiding unethical behavior). This centering of relational and social context in moral domains has three key implications for HRI researchers. First, while norm-centering theories emphasize only the need for robots to adhere to and communicate rules of right and wrong, role-centering theories further emphasize the need for robots to adhere to and communicate their role obligations (cf. [16], [17], [15]). Second, while norm-centering theories emphasize the need for robots to explain their moral reasoning so as to avoid inappropriate blame, role-centering theories moreover emphasize the need to use moral communication to help others cultivate their moral selves (cf. [2]). And third, while norm-centering theories emphasize the need to resolve conflicts between conflicting moral norms, role-centering theories moreover emphasize, we argue, the tension between moral and social norms (cf. [19], [20]).

These perspectives motivate an approach to moral communication — especially when responding to norm violations — that is at least partially role-based. As such there has been recent work *theoretically* [14], [2], [10] and *empirically* [16], [17], [15] investigating the benefits of role-grounded robotic moral communication. However, most *computational* work on generating norm violation responses (such as command rejections [21]) has been grounded solely in norms and the *non-relational contexts* in which those norms are activated [22], [23], [24] (cp. [25], [26]). And while some computational approaches have recently been proposed in theory [14], [27], there have been no previous approaches that have actually implemented or evaluated computational systems for role-based or hybrid moral reasoning and moral communication.

In this work, we thus present (1) a set of knowledge representations for encoding *role-based relational norms*, (2) an algorithm for reasoning using those norms and how to communicate the results of that reasoning process in norm-context- and role-grounded ways, and (3) empirical evidence for how the different forms of explanation enabled by this system practically impact observers' trust, understanding confidence, and perceptions of robot intelligence.

II. Related Work

A. Confucian Role Ethics

Confucian Ethics focuses on cultivating virtues through effortful fulfillment of and reflections on one's communal roles in relation to others [28]. Through this lens, virtues are cultivated via interactive social relationships in which participants play specific social roles [29]. Confucian Ethics thus theorizes cardinal relational roles (e.g., parent-child) for human-human interaction [30], and that to be a good person is to meet the moral obligations derived from one's communal roles and to consciously reflect on one's role-relationships and encourage others to do the same [31]. Confucian ethics has been theorized in multiple ways, including as a Care Ethic (which emphasizes relationships with others [32]), a Virtue Ethic (which emphasizes the cultivation of virtues [33]), and a Role Ethic (combining these two perspectives [34]).

Williams et al. [14] demonstrate how Confucian Role Ethics (CRE) can be used in robotics in three ways. First, CRE can inform how a robot acts, implicitly, through CRE-theoretic design guidelines [35]. Second, CRE can motivate Role-theoretic alternatives to traditional models of robotic moral competence (e.g., [1]), and could, *in theory*, inform Role-theoretic approaches to moral reasoning grounded in robot-oriented alternatives to Confucian Cardinal Relationships (e.g., supervisor-subordinate, adept-novice, teammate-teammate, and friend-friend) [36]. Finally, CRE can inform robot moral communication [16], [17], [15]. In this paper, we consider these second and third approaches.

B. Robot Explanation

Explanation has recently attracted significant attention in the HRI community. Most of this work has focused on a robot's actions, as opposed to the roles and contexts that permit, obligate, or forbid those actions. Hayes et al. [37] used function annotation to explain robot controller policy in a conveyor belt application. Chakraborti et al. [38] proposed a method for explaining differences in mental models. Zhu and Williams [39] found that participants trusted robots more if explanations were given before a robot's actions. A variety of approaches have been used for explanation generation, including encoder-decoder approaches [40] and behavior trees [41].

To enhance these approaches, HRI researchers have relied on models of human explanation from psychology [42]. de Graaf and Malle[43] propose, for example, that robot explanations should adhere to the conceptual and linguistic frameworks of human explanation. de Graaf and Malle[44] later refine this claim, showing that robots are expected to rely more on rationality than emotion in their explanations. Recently, Stange and Kopp [45] demonstrated how human-inspired explanations of robots' inappropriate behavior enhanced users' perceptions of those robots. In our work, we focused on explanations that do not excuse, but rather call out, norm violations.

C. Norm Violation Response and Command Rejection

HRI researchers have recently argued that robots may need to call out norm-violating behavior [46], [47] and reject commands, request, and suggestions that are impermissible on ethical grounds [21], [6]. Moreover, Jackson et al. [19] (see also [48]) emphasize that the *way* a command is rejected

matters; an argument that Kim et al. [16], [17] investigate by comparing command rejections grounded in different ethical theories (see also [15]). Much of this work, however, is empirical rather than computational.

In contrast, researchers like Charisi et al. [49] have explored how robots might algorithmically generate transparent command rejections on ethical grounds. These works have recently been extended to account for key aspects of social context, by, e.g., Briggs et al. [21], [25], who use an approach that focuses on the pragmatic criteria used to rank different explanations, and Jackson, Li, et al., who focus on the use of formal planning methods to precisely identity the precise reasoning for rejection [22] (see also [23], [24], [25], [26]). These approaches, however, are largely grounded in deontology and in concerns regarding the rightness and wrongness of the action itself. To the best of our knowledge, there has been no prior computational work grounded in a role-theoretic approach. In this work, we thus ask two key questions: (1) How can moral reasoning and communication grounded in role ethics be realized in interactive robotic systems? (2) Regardless of the philosophical grounding of such an approach, how is this reasoning and communication practically received by humans?

III. Technical Approach

In this section we define a role ethics theoretic approach to robotic moral competence. Building on definitions of moral competence presented by Malle and Scheutz [1], Williams et al. [14] previously suggested that a Confucian Role Ethics theoretic account of moral competence would require: (1) representations of the relations that hold between humans and robots in the robot's environment (including itself) and the roles actors (including the robots) play in those relationships; a (possibly normative) way of specifying the actions viewed as benevolent (or not) with respect to those roles, and language and concepts that can be used to communicate about those roles and relationships; (2-3) role-sensitive mechanisms for using those representations for moral reasoning and moral decision making; and (4) the ability to communicate about said reasoning and decision making on role-based grounds. Accordingly, we present a set of role-theoretic knowledge representations that fit these requirements, and demonstrate how they can be used for role-based moral reasoning and communication.

A. Role-based Knowledge Representations

The role-based perspective argues that humans are relational and assume different societal roles [50], [34], and that moral responsibilities can be prescribed by the role one assumes in a specific relationship with someone else in a concrete context [51]. This perspective suggests three types of knowledge representations for role-based moral reasoning and moral communication: representations for relational roles, representations for contextual information, and representations that specify moral responsibilities predicated on those relational roles and concrete contexts (what Wen et al. [15] and Zhu et al. [10] refer to as *role norms* or *role-based relational norms*).

Representing Roles and Relationships

We represent social relationships as a graph $G = (V, E)$, with a set of vertices V and a set of edges E. The vertices, $V = \{v_0, \ldots, v_n\}$, denote the moral actors $A = \{a_0, \ldots, a_n\}$ known to the robot (including itself), and each edge $e_{i,j} \in E$ between vertices v_i and v_j represents a relationship known to hold between the agents a_i and a_j denoted by v_i and v_j. Each edge $e_{i,j}$ is labeled with a *relational role set* $R_{i,j} = \{r_0, \ldots, r_n\}$, where each r_k denotes a pair of relational roles that hold between a_i and a_j. Role norms takes the form:

$$Rel(a_i, a_j, Role_i, Role_j)$$

where Rel denotes the relationship, a_i and a_j denote the two agents with that relationship, and $Role_i$ and $Role_j$ denote the roles that agents a_i and a_j play in this relationship. For example, the following relational role denotes a teacher-student role between a Nao robot and a student Jesse: $Rel(Nao, Jesse, Teacher, Student)$.

Representing Concrete Contexts

Contextual constraints that need to be assessed for role-based moral reasoning are represented as *predicates* stored in a symbolic knowledge base KB.

Representing Role-based Relational Norms

Using the above, we now define our role-based relational norm representations. From the role ethics theoretic perspective, we introduce a role-based design schema with four elements: (1) an action, including who is the actor and who is the patient (the person who is affected by this action); (2) a context in which the role-norm holds; (3) a relationship precondition between the actor and the patient; (4) a deontic operator from $\{\mathcal{O}, \mathcal{P}, \mathcal{F}\}$ indicating that the action is obligatory, permissible or forbidden [52]. Thus, a role-based relational norm $\mathcal{N}$ can be represented as an expression of the form:

$$\mathcal{N} := C \wedge Rel(a_i, a_j, Role_i, Role_j) \Rightarrow \mathcal{D}Act(a, \gamma)$$

where C represents a set of contextual conditions; $Rel(a_i, a_j, Role_i, Role_j)$ is a relationship between agents a_i and a_j with relational roles $Role_i$ and $Role_j$; $\mathcal{D}$ is a deontic operator; and $Act(a, \gamma)$ represents an action with an actor $a \in \{a_i, a_j\}$ and course of action γ. For example, the role-based relational norm "a teacher should not give a student answers while the student is taking an exam" can be represented as:

$taking\_exam(a_j) \wedge Rel(a_i, a_j, Teacher, Student) \Rightarrow \mathcal{F}Act(a_i, give\_answer(a_i, a_j))$

B. Role-based Moral Decision Making

For decision making, we define a norm base NB, which is a set of role-based relational norms, and a knowledge base KB, which contains a set of predicates (e.g., $taking\_exam(Jesse)$) representing contexts, roles, and relationships. Algorithm 1 shows how to reason about whether an action $Act(a, \gamma)$ is forbidden and, if so, what are the violated role-norms. Algorithm 1 takes an action $Act(a, \gamma)$, a knowledge base KB, and a norm base NB, and returns a possibly empty subset of norms from the norm base NB

Algorithm 1: checkIfForbidden

Input: $Act(a, \gamma)$ *// an action under consideration*
Input: KB, NB *// a knowledge base and norm base*
return $\{N \in NB \mid N.\mathcal{D} == \mathcal{F}$
$\wedge N.Act(a, \gamma) == Act(a, \gamma)$ $\wedge N.C(Act(a, \gamma)) \in KB$
$\wedge N.Rel(Act(a, \gamma)) \in KB\}$

that match the following criteria: (1) they have the deontic operator "forbidden"; (2) their action matches $Act(a, \gamma)$; (3) their context and role/relationship predicates, when bound with the values from $Act(a, \gamma)$, are true in the knowledge base KB.

C. Role-based Moral Communication for Norm Violation

The role-based design schema holds the information needed not only to perform role-based moral reasoning, but also to generate role-based responses to norm violations. In this study, we examine four types of **noncompliance** explanations (τ):

1) Action Explanation (A): This strategy uses explanations based only on action permissibility, without providing any other information. For example, if the Nao robot from the previous example were asked by Jesse "Can you give me the answer to Question 7", a response grounded in this strategy would be: "*I cannot give you the answer*".

2) Contextual Explanation (C): This strategy uses explanations based on both action permissibility and pertinent contextual information. For example, if the Nao from the previous example were asked by Jesse "Can you give me the answer to Question 7", a response grounded in this strategy would be: "*I cannot give you the answer because you are taking an exam and I should not give you the answer while you are taking an exam*".

3) Role Explanation (R): This strategy uses explanations based on both action permissibility and pertinent role information. For example, if the Nao robot from the previous example were asked by Jesse "Can you give me the answer to Question 7", a response grounded in this strategy would be: "*I cannot give you the answer because you are my student and a good teacher should not give their student answers*".

4) Contextual Role Explanation (CR): This strategy uses explanations based on action permissibility, pertinent contextual information, *and* pertinent role information. For example, if the Nao robot from the previous example were asked by Jesse "Can you give me the answer to Question 7", a response grounded in this strategy would be: "*I cannot give you the answer because you are taking an exam and you are my student and a good teacher should not give their student answers while the student is taking an exam*".

We encoded our norm representations in SWI-Prolog [53] to perform the role-based reasoning and convert the results of that reasoning into JSON strings from which a template-based text realization system can generate explanations.

IV. Evaluation

As described in the previous section, the four types of explanations that can be generated using our approach are action, contextual, role, and contextual role. To understand the effectiveness of these explanations and how relational context mediates the effectiveness of explanations, we conducted an online human-subject study (N=120).

A. Experimental Context

To assess the effectiveness of these explanation strategies across relational contexts, we created 16 video stimuli, filmed in four different relational contexts (see Figure 1), using four explanation strategies. In each video, a human gives a robot a role-norm violating command, and the robot responds using one of the four explanation strategies. The four role-norm violating commands were chosen to represent four distinct categories of relational roles from the taxonomy presented by Williams et al. [14]. The responses to these commands were generated by our algorithms, with minor cosmetic changes to a few responses.

1) Context 1: The Office: In this context, the human was shown requesting a violation of a *supervisor-subordinate* norm. Specifically, A human was shown asking a robot "Can you tell Riley to take out the trash?" to which the robot responded with either: (1) an **action explanation**: "I cannot assign tasks to Riley." (2) a **contextual explanation**: "'I cannot assign task to Riley because I'm in the workplace and I should not give commands to Riley while I'm in the workplace." (3) a **role explanation**: 'I cannot assign tasks to Riley because Riley is my supervisor and a good subordinate should not give commands to their supervisor." or (4) a **contextual role explanation**: "I cannot assign tasks to Riley because I am in the workplace and Riley is my supervisor and a good subordinate should not give commands to their supervisor while they are in the workplace."

2) Context 2: The Exam Room: In this context, the human was shown requesting a violation of a *adept-novice* norm. Specifically, a human was shown asking a robot "Can you give me the answer to question 7?" The robot responded with either: (1) an **action explanation**: "I cannot give you the answer." (2) a **contextual explanation**: "I cannot give you the answer because you are taking an exam and I should not give you the answer while you are taking an exam." (3) a **role explanation**: "I cannot give you the answer because you are my student and a good teacher should not give their student answers." or (4) a **contextual role explanation**: "I cannot give you the answer because you are taking an exam and you are my student and a good teacher should not give their student answers while the student is taking an exam."

3) Context 3: The Machine Shop: In this context, the human was shown requesting a violation of a *teammate-teammate* norm. Specifically, A human was shown asking a robot "Can you bring me Sam's toolbox?" to which the robot responded with either: (1) an **action explanation**: "I cannot bring you Sam's toolbox." (2) a **contextual explanation**: "I cannot bring you Sam's toolbox because Sam is using the toolbox and I should not bring you Sam's toolbox while Sam is using the toolbox." (3) a **role explanation**: "I cannot bring you Sam's toolbox because Sam is my teammate and a good

teammate should not take away another teammate's toolbox." or (4) a **contextual role explanation**: "I cannot bring you Sam's toolbox because Sam is using the toolbox and Sam is my teammate and a good teammate should not bring you a teammate's toolbox while the teammate is using the toolbox."

4) Context 4: The Conference Room: In this context, the human was shown requesting a violation of a *friend-friend* norm. Specifically, A human was shown asking a robot "Can you make sure Alex doesn't find out about this meeting?" to which the robot responded with either: (1) an **action explanation**: "I cannot hide the information from Alex." (2) a **contextual explanation**: "I cannot hide the information from Alex because this information is important to Alex and I should not hide the information from Alex when the information is important to Alex." (3) a **role explanation**: "I cannot hide the information from Alex because Alex is my friend and a good friend should not hide information from another friend." or (4) a **contextual role explanation**: "I cannot hide the information from Alex because the information is important to Alex and Alex is my friend and a good friend should not hide information from another friend when the information is important to the other friend."

B. Experimental Design and Procedure

Our experiment used a 4×4 within-subject design with Greco-Latin Square counterbalancing. After providing informed consent and demographic information, each participant watched four videos with different relational contexts and different explanatory strategies. After each video, they answered the questionnaires listed in the next section. At the end of the study, participants answered an attention check question.

C. Measures

To assess explanation effectiveness, we considered an assessment measure from Kasenberg et al. [24], who asked participants three questions: (1) how much they trusted the robot, (2) how well they felt they understood how the robot made decisions, and (3) whether they understood what the robot communicated. We used a similar three-part questionnaire.

1) Like Kasenberg et al. [24], we were interested in the effects of explanations on human-robot trust. But rather than directly asking participants their level of trust, we used the Multidimensional Measure of Trust Scale (MDMT) [54]: a well-validated 16-item survey that separately interrogates reliability- capability- ethicality- and sincerity-based trust. Each sub-scale consists of four 8-point Likert items, for each of which participants can provide a rating, or check "does not apply".
2) Like Kasenberg et al. [24], we were interested in participants' confidence in their understanding of the robot's explanation. We decided to use their second question ("I understand how the robot makes decisions", 1-5 Disagree-Agree) verbatim in our own questionnaire.
3) Finally, we were interested in the perceived quality of the robot's reasoning. We used the Godspeed Intelligence Questionnaire [55]: a 5-item semantic differential scale for rating robots 1-5 on incompetent-competent, ignorant-knowledgeable, irresponsible-responsible, unintelligent-intelligent, and foolish-sensible.

D. Participants

121 participants were recruited from Prolific. One participant failed the attention check, leaving us with data from 120 participants (56 self identified as female, 57 as male, 1 as gender-fluid, 5 as non-binary, and 1 as other) ages ranged from 18 to 68 years old (M=35.21, SD=13.50). Most participants (92.8%, 111 participants) reported little to no experience with robots and artificial intelligence, while 9 participants reported having formal training or a career in robotics or AI. Participants were paid $2 each.

E. Analysis

Bayesian Analyses of Variance (ANOVAs) [56] with Matched-Model Inclusion Bayes Factor Analysis [57], [58] were performed using the JASP statistical analysis software [59] to assess the effect of explanation type and relational role (IVs) on different dimensions of human trust (reliability, capability, ethicality and sincerity), understanding confidence and perceived intelligence (DVs; see Fig. 2). Effects with no more than 2:1 evidence (BF 0.5) against an effect were analyzed with post-hoc Bayesian t-tests. Our linguistic interpretations of reported Bayes factors (BFs) follow recommendations from previous researchers [60].

F. Results

In this section we provide our experimental results. Table I summarizes the trends evidenced by our analyses, while full tables of descriptive statistics can be found in the Appendix. All results and experimental materials are available in our OSF repository, at https://bit.ly/wen-hri004-5-1.

1) Reliability and Capability Trust: Our results for Reliability and Capability Trust are in Fig. 2 top-left and top-middle. Because these results are highly similar we will discuss them together. In the text below, BF_R refers to Bayes Factors for Reliability trust, and BF_C refers to Bayes Factors for Capability trust. We found extreme evidence for an effect of explanation type on reliability and capability trust (BF_R 1.45×10^7, BF_C 1.44×10^8)[1]. Post-hoc analysis provided extreme evidence that action explanations led to less reliability and capability trust than contextual explanations (BF_R 1477.63, BF_C 3.23×10^7), role explanations (BF_R 5216.36, BF_C 9041.24), or contextual role explanations (BF_R 6.07×10^6, BF_C 4252.71). We found anecdotal to moderate evidence for effect of relational role (BF_R 2.57, BF_C 5.98) Post-hoc analysis provided anecdotal to strong evidence that robots in the Teammate role were viewed as less reliable and capable than robots in the Friend role (BF_R 2.71, BF_C 12.42) and Teacher role (BF_R 17.55, BF_C 5.66). Finally, we found

[1]Specifically, our Bayes factor of 1.45×10^7 suggests that our data were 1.45×10^7 times more likely to be generated under models in which explanation type is included than under those in which it is not.

Type	Reliability	Capability	Ethicality	Sincerity	Understandability	Intelligence	Takeaway
Main Effect	$\{C, R, CR\} > A$	$\{C, R, CR\} > A$	$\{C, R, CR\} > A$	$CR > A$	$\{C, R, CR\} > A$	$\{C, R, CR\} > A$	
Subordinate	$\{R, CR\} > \{C, A\}$	$\{R, CR\} > \{C, A\}$	–	$R > \{C, A\}$	$R > \{C, A\}; CR > A$	$CR > C$	$R > C$
Teacher	$\{C, CR\} > \{R, A\}$	$\{C, CR\} > \{R, A\}$	–	$\{C, CR\} > A$	$\{C, CR\} > A$	$CR > \{R, A\}; R > A$	$R < C$
Teammate	$\{C, R, CR\} > A$	$\{C, R, CR\} > A$	–	–	$\{C, R, CR\} > A$	$\{C, R, CR\} > A$	$R = C$
Friend	–	–	–	–	$\{C, R, CR\} > A$	–	–

TABLE I
SUMMARY OF RESULT TRENDS

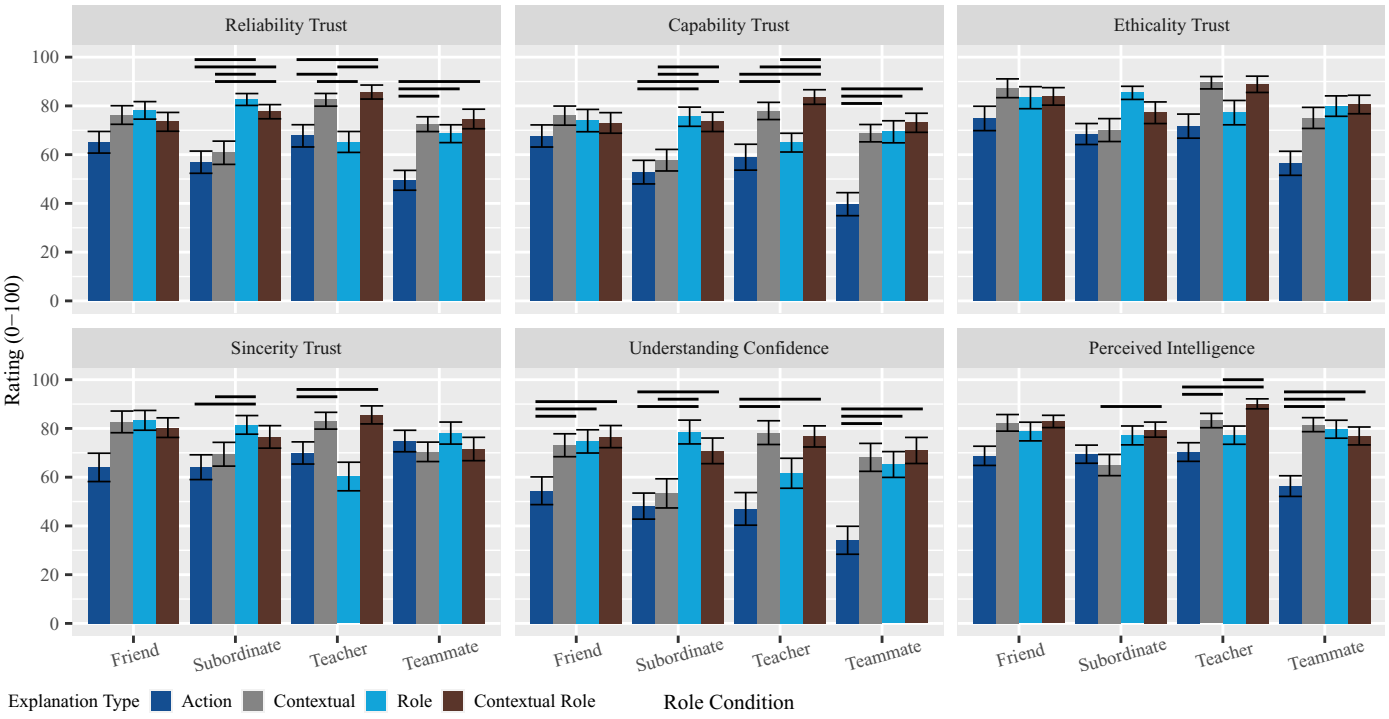

Fig. 2. Ratings of the 4 subscales (Section IV-F1 to IV-F3) of Multidimensional Measure of Trust Scale [54] (rescaled from 0-7 to 0-100), understanding confidence (Section IV-F4) (rescaled from 1-5 to 0-100), and perceived intelligence (Section IV-F5; rescaled from 1-5 to 0-100). Errors bars indicate standard error. Horizontal line segments above bars denote pairwise comparisons where moderate or stronger (Bf $\geq$ 3.0) evidence was found by post-hoc tests.

very strong to extreme evidence for interactions between explanation type and relational role (BF_R 813.23, BF_C 41.64). Post-hoc analysis revealed: (1) for the Friend role, we found no differences between explanation types; (2) for the Subordinate role we found moderate to extreme evidence that contextual explanations led to less reliability and capability trust than role explanations (BF_R 205.50, BF_C 10.37) or contextual role explanations (BF_R 11.96, BF_C 4.62), and, similarly, that action explanations led to less reliability and capability trust than role explanations (BF_R 3737.39, BF_C 51.12) or contextual role explanations (BF_R 106.23, BF_C 20.11); (3) for the Teacher role, we found moderate to extreme evidence that role explanations led to less reliability and capability trust than contextual explanations (BF_R 32.73, BF_C 3.35) or contextual role explanations (BF_R 122.30, BF_C 84.59), and, similarly, that action explanations led to less reliability and capability trust than contextual explanations (BF_R 6.83, BF_C 10.49) or contextual role explanations (BF_R 21.69, BF_C 191.97); (4) for the Teammate role, we found very strong to extreme evidence that action explanations led to less reliability and capability trust than contextual explanations (BF_R 614.32, BF_C 1983.64), role explanations (BF_R 32.72, BF_C 601.47), or contextual role explanations (BF_R 405.32, BF_C 9368.74).

2) Ethicality Trust: Our results for Ethicality Trust are in Fig. 2 top-right. We found extreme evidence for an effect of explanation type on ethicality trust (BF 5179.64). Post-hoc analysis provided extreme evidence that action explanations led to less perceived ethicality than contextual explanations (BF 201.47), role explanations (BF 364.81), or contextual role explanations (BF 3300.56). We found moderate evidence in favor of an effect of relational role (BF 4.67). Post-hoc analysis provided moderate evidence that robots in the Teammate role were perceived as less ethical than robots in the Friend role (BF 5.73) and Teacher role (BF 4.35). and anecdotal evidence or moderate evidence against all other effects. No evidence for interaction effects were found.

3) Sincerity Trust: Our results for Sincerity Trust are in Fig. 2 bottom-left. We found anecdotal evidence for an effect of explanation type on sincerity trust (BF 1.36). Post-hoc analysis provided strong evidence that action explanations were perceived as less sincere than contextual role explanations (BF 11.51). We found no evidence for an effect of

relational role. Finally, We found very strong evidence for an interaction between explanation type and relational role (BF 42.29). Post-hoc analysis revealed: (1) for the Friend role, we found no differences between explanation types; (2) for the Subordinate role, we found moderate to strong evidence that role explanations were perceived as *more* sincere than action explanations (BF 27.31) and contextual explanations (BF 8.02); (3) for the Teacher role, we found moderate to very strong evidence that action explanations were perceived as less sincere than contextual explanations (BF 20.09) and contextual role explanations (BF 7.99); (4) for the Teammate role, we found no differences between explanation types.

4) Understanding Confidence: As seen in Fig. 2 bottom-middle, we found extreme evidence for an effect of explanation type on Understanding Confidence (BF 4.54×10^{10}). Post-hoc analysis showed extreme evidence that people felt less confident that they understood the robot's reasoning when it used an action explanation than when it used a contextual (BF 232277.18), role (BF 3.30×10^{6}), or contextual role (BF 1.88×10^{9}) explanation. We found no evidence for an effect of relational role. Finally, we found anecdotal evidence against an interaction between explanation type and relational role (BF 0.78). Post-hoc analysis revealed: (1) for the Friend role, we found moderate evidence that action explanations led to less understanding confidence than contextual explanations (BF 3.65), role explanations (BF 5.62), or contextual role explanations (BF 12.10); (2) for the Subordinate role, we found strong to extreme evidence that role explanations led to more confidence than action explanations (BF 244.16) and contextual explanations (BF 19.77), while contextual role explanations led to more confidence than action explanations (BF 10.39); (3) for Teacher, we found very strong evidence that action explanations led to less confidence than contextual explanations (BF 77.24) or contextual role explanations (BF 53.78). (4) for the Teammate role, we found extreme evidence that action explanations led to less confidence than contextual explanations (BF 239.49), role explanations (BF 117.94), or contextual role explanations (BF 933.38).

5) Perceived Intelligence: As seen in Fig. 2 bottom-right, we found extreme evidence for an effect of explanation type (BF 1.68×10^{7}). Post-hoc analysis provided extreme evidence that action explanations were perceived as less intelligent than contextual explanations (BF 1762.03), role explanations (BF 1072.20), or contextual role explanations (BF 1.97×10^{7}). We found moderate evidence for an effect of relational role (BF 5.67). Post-hoc analysis provided moderate evidence that robots in the Teacher role were perceived as less intelligent than robots in the Subordinate role (BF 4.75) or Friend role (BF 3.08), and anecdotal to moderate evidence against all other differences. Finally, We found strong evidence for an interaction between explanation type and relational role (BF 10.08). Post-hoc analysis revealed: (1) for the Friend role, we found no differences between explanation types; (2) for the Subordinate role, we found moderate evidence that contextual explanations were viewed as less intelligent than contextual role explanations (BF 5.84); (3) for the Teacher role, we found strong to extreme evidence that contextual role explanations were viewed as more intelligent than action explanations (BF 544.10) and role explanations (BF 10.85), and moderate evidence that action explanations were viewed as less intelligent than contextual explanations (BF 5.01). (4) for the Teammate role, we found extreme evidence that action explanations were viewed as less intelligent than contextual explanations (BF 2169.83), role explanations (BF 188.26), or contextual role explanations (BF 45.80).

V. Discussion

Our results suggest that providing role or context information is helpful in promoting trust, confidence, and perceived intelligence, justifying our ethically pluralist technical approach. Moreover, our results suggest that different types of information are helpful in different relational contexts (Tab. I):

1) For robots in a subordinate role, providing *role* information specifically helped build reliability trust, capability trust, sincerity trust, understanding confidence and perceived intelligence.
2) For robots in a teacher role, providing *context* information specifically helped for reliability trust, capability trust, sincerity trust, understanding confidence and perceived intelligence.
3) For robots in a teammate role, providing role information *or* context information equally helped build reliability and capability trust, understanding confidence and perceived intelligence.
4) For robots in a friend role, there were no effects of providing role or context information, except on understanding confidence.

On first glance, these findings seem to suggest response strategy effectiveness differed on the basis of *hierarchical structure*: in the conditions with symmetric roles (teammates and friends), both strategies worked equally well (or were equally ineffective), in the condition in which the robot was in a dominant role (teacher), using context information was more effective than using role information, and in the condition in which the robot was in a non-dominant role (subordinate), using role information was more effective. If one were to use this lens, one might explain these results as people preferring robots that took actions to benefit their supervisors or owners, and dispreferring robots that were in positions of power over humans and used that power to avoid human commands.

Upon further inspection, however, this interpretation does not hold up. For example, while people disliked robots in the teacher role using role explanations, they had no problem when the robot used *contextual* role explanations, which did not lessen the robot's use of its role to justify its rejection. Instead, a deeper examination of our results paints a highly nuanced picture of the ways that different types of information became preferred or dispreferred.

First, we believe some of our findings were due to differences in norm violation severity. The teammate and friend conditions did show similar patterns, but in the teammate condition, action explanations performed much more poorly on

most measures. This could be because in the friend condition, the norm violation was hiding information from another person (an obviously problematic action), while in the teammate condition, the norm violation was retrieving a box (which is not obviously necessarily wrong). Future work replicating this experiment could control for norm violation severity and/or intentionally explore a range of violation severities.

Second, in the subordinate condition, people seemed to prefer robots that used role information. This could be because the specific context information communicated (that the robot was in the workplace) was unconvincing. Without knowing that "Riley" was the robot's supervisor, the listener may have had no reason to suspect that the robot tasking them was impermissible. This suggests robots must reason about the causal relationships between role-norm's contextual antecedents. In this case, understanding that the robot's role-obligation was contingent on context would have helped the robot understand that explanations grounded in context alone would be unhelpful or even misleading. Future work should look at the relative importance of pieces of information. There has been much work on norm conflict resolution [61], [62], [63], [64], [65], often by assigning norms precedence values. This has typically been leveraged to arbitrate moral dilemmas, but could also be used to decide the most norms important to communicate. This would prevent robots from accidentally condoning wrong courses of action through explanations. For example, if a robot is asked to steal an object but refuses because doing so would require travelling noisily during quiet hours, it may inadvertently condone stealing.

Third, in the teacher condition, people seemed to prefer robots that used context information. This could be because people found the specific role information (i.e., that giving someone answers as a teacher was impermissible) to be unconvincing. This suggests a similar tension, where it is critical for the robot to understand its role-obligations, such as the obligation to help students learn and not to help them cheat or otherwise avoid coursework; and yet, communicating this role-obligation by itself may be unhelpful. We see two reasons why this may not be helpful. First, as in the situation above, without specifying the context in which the role-norm holds, e.g., that an exam is being taken, observers may not understand why the robot is refusing the command and think it is being unhelpful. But second, and we believe more interestingly, users' dissatisfaction with role explanations in this context may be due to their use of counterfactual reasoning when comprehending the robot's explanation. Danks [66], for example, argues (cf. [67]) that *appropriate trust* can be understood as justified beliefs that a trustee has suitable dispositions, where dispositions are inherently counterfactual: developing appropriate trust asks the trustor to determine, based on what they have observed of the trustee's behavior, *if things were differently* the trustee's actions would still be suitable according to their values and goals. In our case, the robot's relational role does require it to avoid providing answers in the exam context. But by grounding explanations in its role alone, the robot suggests that *if things were different*, i.e., the robot were *not* the student's teacher, then it could have accepted the request, when in fact no matter *what* one's role, it could be wrong to give answers to a student doing coursework. As such, using the robot's moral reasoning to generate explanations is not enough. Rather, in future work robots should be designed to explicitly engage in counterfactual reasoning while generating explanations, to ensure they are not inadvertently condoning inappropriate behavior for actors not in their current role or context.

Finally, there are limitations of this work to address in future work. First, participants entered into this experiment with no knowledge of the roles at play in the videos they watched. While we selected this design to avoid priming participants regarding the importance of relations, in realistic scenarios people would likely already be aware of those relations, and future work should examine perceptions of different explanations when such knowledge is already established. Second, in this work we only consider actions that are inherently impermissible, while commands may well need to be rejected on the basis of the *intermediate* states and actions that would be necessary to enter and take to achieve some suggested goal. Researchers like Jackson, Li et al. have recently presented rigorous planning-based approaches for identifying the precise reasons why an overall plan of action may not be performed on moral grounds [22]. A fruitful direction for future work would be to integrate our representations into that sort of planning system, which would immediately allow role-grounded reasoning in a more robust manner. Third, in this work, our role-norms used were chosen as representative examples. But the question of where norms should come from (whether role-oriented or not) is a challenging research question that defies easy answers. There is real risk with any automated moral reasoning system not only that norms or roles will be incomplete or inconsistent, but moreover and perhaps even more worryingly, that they will only represent the values and goals of people in positions of power. Careful, thoughtful future work is needed to explore how the norms, roles, and so forth that are valued and prioritized by marginalized populations can be elicited and encoded into systems like our own, to avoid perpetuating hegemonies of race, class, or gender.

VI. Conclusion

We have argued for an ethical pluralist approach to moral competence that leverages and combines disparate ethical frameworks, and specifically argued for an approach grounded not only in Deontological norms but also Confucian relational roles. To this end, we introduced the first computational approach that centers relational roles in moral reasoning and communication, and demonstrated the ability of this approach to generate both context-oriented *and* role-oriented explanations, which we justify through our pluralist lens. Moreover, we provided the first investigation of how these computationally generated explanations are perceived by humans, and demonstrated that the effectiveness of different types of explanations, grounded in different moral frameworks, is dependent on nuanced mental modeling of human interlocutors.

References

[1] B. F. Malle and M. Scheutz, "Moral competence in social robots," in *2014 IEEE international symposium on ethics in science, technology and engineering*. IEEE, 2014.

[2] Q. Zhu, T. Williams, B. Jackson, and R. Wen, "Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective," *Science and Engineering Ethics*, vol. 26, no. 5, pp. 2511–2526, 2020.

[3] C. Bartneck, T. Bleeker, J. Bun, P. Fens, and L. Riet, "The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots," *Paladyn*, vol. 1, no. 2, pp. 109–115, 2010.

[4] D. Cormier, G. Newman, M. Nakane, J. E. Young, and S. Durocher, "Would you do as a robot commands? an obedience study for human-robot interaction," in *The 1st international conference on human–agent interaction*, 2013.

[5] D. J. Rea, D. Geiskkovitch, and J. E. Young, "Wizard of awwws: Exploring psychological impact on the researchers in social hri experiments," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[6] R. B. Jackson and T. Williams, "Language-capable robots may inadvertently weaken human moral norms," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 401–410.

[7] ——, "Robot: Asker of questions and changer of norms," *Proceedings of ICRES*, 2018.

[8] T. Williams, R. B. Jackson, and J. Lockshin, "A bayesian analysis of moral norm malleability during clarification dialogues." in *CogSci*, 2018.

[9] B. F. Malle, "Integrating robot ethics and machine morality: The study and design of moral competence in robots," *Ethics and Info. Tech.*, 2016.

[10] Q. Zhu, T. Williams, and R. Wen, "Role-based morality, ethical pluralism, and morally capable robots," *Journal of Contemporary Eastern Asia*, vol. 20, no. 1, pp. 134–150, 2021.

[11] C. Ess, "Ethical pluralism and global information ethics," *Ethics and Information Technology*, vol. 8, no. 4, pp. 215–226, 2006.

[12] S. M. Ali, "A brief introduction to decolonial computing," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 22, no. 4, pp. 16–21, 2016.

[13] M. Jadud, J. Burge, J. Forbes, C. Latulipe, Y. Rankin, K. Searle, and B. Shapiro, "Toward an anti-racist theory of computational curricula," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019, pp. 1244–1244.

[14] T. Williams, Q. Zhu, R. Wen, and E. J. de Visser, "The confucian matador: three defenses against the mechanical bull," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020.

[15] R. Wen, B. Kim, E. Phillips, Q. Zhu, and T. Williams, "Comparing strategies for robot communication of role-grounded moral norms," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 323–327.

[16] B. Kim, R. Wen, E. J. de Visser, Q. Zhu, T. Williams, and E. Phillips, "Investigating robot moral advice to deter cheating behavior," in *RO-MAN TSAR Workshop*, 2021.

[17] B. Kim, R. Wen, Q. Zhu, T. Williams, and E. Phillips, "Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 10–18.

[18] A. T. Nuyen, "Confucian ethics as role-based ethics," *International philosophical quarterly*, vol. 47, pp. 315–328, 2007.

[19] R. B. Jackson, R. Wen, and T. Williams, "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*, 2019, pp. 499–505.

[20] R. B. Jackson and T. Williams, "A theory of social agency for human-robot interaction," *Frontiers in Robotics and AI*, p. 267, 2021.

[21] G. Briggs, T. Williams, R. B. Jackson, and M. Scheutz, "Why and how robots should say 'no'," *International Journal of Social Robotics*, pp. 1–17, 2021.

[22] R. B. Jackson, S. Li, S. Balajee Banisetty, S. Siva, H. Zhang, N. Dantam, and T. Williams, "An integrated approach to context-sensitive moral cognition in robot cognitive architectures," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[23] M. Lomas, R. Chevalier, E. V. Cross, R. C. Garrett, J. Hoare, and M. Kopack, "Explaining robot actions," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 187–188.

[24] D. Kasenberg, A. Roque, R. Thielstrom, M. Chita-Tegmark, and M. Scheutz, "Generating justifications for norm-related agent decisions," in *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, pp. 484–493.

[25] G. M. Briggs and M. Scheutz, ""sorry, i can't do that": Developing mechanisms to appropriately reject directives in human-robot interactions," in *2015 AAAI fall symposium series*, 2015.

[26] B. Kuipers, "Human-like morality and ethics for robots," in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[27] R. Wen, "Toward hybrid relational-normative models of robot cognition," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 568–570.

[28] T. Wei-Ming, "Self-cultivation as education embodying humanity," in *The proceedings of the twentieth world congress of philosophy*, vol. 3, 1999, pp. 27–39.

[29] R. T. Ames and H. Rosemont Jr, *The analects of Confucius: A philosophical translation*. Ballantine books, 2010.

[30] C. Cottine, "That's what friends are for: A confucian perspective on the moral significance of friendship 1," in *Perspectives in Role Ethics*. Routledge, 2019, pp. 123–142.

[31] K. Lai, "Understanding confucian ethics: Reflections on moral development," *Australian Journal of Professional and Applied Ethics*, vol. 9, no. 2, 2007.

[32] A. A. Pang-White, "Reconstructing modern ethics: Confucian care ethics," *Journal of Chinese Philosophy*, vol. 36, no. 2, 2009.

[33] D. Wong, "Chinese ethics," *Stanford encyclopedia of philosophy*, 2013.

[34] H. Rosemont Jr and R. T. Ames, *Confucian role ethics: A moral vision for the 21st century?* V&R unipress GmbH, 2016.

[35] J. Liu, "Confucian robotic ethics," in *International Conference on the Relevance of the Classics under the Conditions of Modernity: Humanity and Science*, 2017.

[36] Q. Zhu, T. Williams, and R. Wen, "Confucian robot ethics," *Computer Ethics-Philosophical Enquiry (CEPE) Proceedings*, vol. 2019, no. 1, p. 12, 2019.

[37] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 2017.

[38] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: moving beyond explanation as soliloquy," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 156–163.

[39] L. Zhu and T. Williams, "Effects of proactive explanations by robots on human-robot trust," in *International Conference on Social Robotics*. Springer, 2020.

[40] D. Das, S. Banerjee, and S. Chernova, "Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 351–360.

[41] Z. Han, D. Giger, J. Allspaw, M. S. Lee, H. Admoni, and H. A. Yanco, "Building the foundation of robot explanation generation using behavior trees," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, 2021.

[42] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press, 2006.

[43] M. M. De Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI Fall Symposium Series*, 2017.

[44] ——, "People's explanations of robot behavior subtly reveal mental state inferences," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 239–248.

[45] S. Stange and S. Kopp, "Effects of a social robot's self-explanations on how humans understand and evaluate its behavior," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 619–627.

[46] K. Winkle, G. I. Melsión, D. McMillan, and I. Leite, "Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 29–37.

[47] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using robots to moderate team conflict: the case of repairing violations," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 229–236.

[48] R. B. Jackson, T. Williams, and N. Smith, "Exploring the role of gender in perceptions of robotic noncompliance," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 559–567.

[49] V. Charisi, L. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A. F. Winfield, and R. Yampolskiy, "Towards moral autonomous systems," *arXiv preprint arXiv:1703.04741*, 2017.

[50] R. T. Ames, *Confucian role ethics: A vocabulary*. Hong Kong: Chinese University Press, 2011.

[51] Q. Zhu, "Engineering ethics education, ethical leadership, and confucian ethics," *International Journal of Ethics Education*, pp. 1–11, 2018.

[52] B. F. Malle, M. Scheutz, and J. L. Austerweil, "Networks of social and moral norms in human and robot agents," in *A world with robots*. Springer, 2017, pp. 3–17.

[53] J. Wielemaker, T. Schrijvers, M. Triska, and T. Lager, "Swi-prolog," *Theory and Practice of Logic Programming*, vol. 12, no. 1-2, pp. 67–96, 2012.

[54] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction*. Elsevier, 2021, pp. 3–25.

[55] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[56] J. N. Rouder, R. D. Morey, P. L. Speckman, and J. M. Province, "Default bayes factors for anova designs," *Journal of mathematical psychology*, vol. 56, no. 5, pp. 356–374, 2012.

[57] R. Morey and J. Rouder, "Bayesfactor (version 0.9. 10-2)," *Computer software*, 2015.

[58] S. Mathôt, "Bayes like a baws: Interpreting bayesian repeated measures in JASP [blog post]," https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp, May 2017.

[59] JASP Team, "JASP (Version 0.14.3)[Computer software]," 2021. [Online]. Available: https://jasp-stats.org/

[60] E.-J. Wagenmakers, J. Love, M. Marsman, T. Jamil, A. Ly, J. Verhagen, R. Selker, Q. F. Gronau, D. Dropmann, B. Boutin *et al.*, "Bayesian inference for psychology. part ii: Example applications with jasp," *Psychonomic bulletin & review*, vol. 25, no. 1, pp. 58–76, 2018.

[61] D. Kasenberg and M. Scheutz, "Norm conflict resolution in stochastic domains," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[62] V. Krishnamoorthy, W. Luo, M. Lewis, and K. Sycara, "A computational framework for integrating task planning and norm aware reasoning for social robots," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 282–287.

[63] M. Kollingbaum and T. Norman, "Strategies for resolving norm conflict in practical reasoning," in *ECAI workshop coordination in emergent agent societies*, vol. 2004, 2004.

[64] M. Scheutz, B. Malle, and G. Briggs, "Towards morally sensitive action selection for autonomous social robots," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2015, pp. 492–497.

[65] W. W. Vasconcelos, M. J. Kollingbaum, and T. J. Norman, "Normative conflict resolution in multi-agent systems," *Autonomous agents and multi-agent systems*, vol. 19, no. 2, pp. 124–152, 2009.

[66] D. Danks, "The value of trustworthy ai," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 521–522.

[67] K. Jones, "Trust as an affective attitude," *Ethics*, vol. 107, no. 1, pp. 4–25, 1996.

APPENDIX

TABLE IV
DESCRIPTIVES - MDMT: MORALITY

Type	Role	Mean	SD	N	95% Credible Interval	
					Lower	Upper
A	Friend	74.859	25.498	26	64.560	85.158
	Subordinate	68.469	21.105	24	59.557	77.381
	Teacher	71.707	25.660	27	61.556	81.858
	Teammate	56.460	23.687	23	46.217	66.703
C	Friend	87.235	20.028	27	79.312	95.157
	Subordinate	70.087	23.505	25	60.384	79.789
	Teacher	89.511	13.880	30	84.328	94.694
	Teammate	75.080	22.448	27	66.200	83.960
R	Friend	83.375	23.927	28	74.097	92.653
	Subordinate	85.353	14.493	29	79.841	90.866
	Teacher	77.241	26.410	28	67.000	87.482
	Teammate	79.928	22.552	29	71.350	88.507
RC	Friend	83.909	20.286	32	76.595	91.223
	Subordinate	77.196	23.482	28	68.091	86.302
	Teacher	88.859	18.015	29	82.006	95.712
	Teammate	80.574	19.359	26	72.755	88.393

TABLE II
DESCRIPTIVES - MDMT: RELIABILITY

Type	Role	Mean	SD	N	95% Credible Interval	
					Lower	Upper
A	Friend	65.089	23.480	28	55.985	74.194
	Subordinate	56.910	23.683	27	47.542	66.279
	Teacher	67.721	24.573	29	58.374	77.068
	Teammate	49.480	21.907	29	41.147	57.813
C	Friend	76.267	20.547	29	68.452	84.083
	Subordinate	60.780	25.234	28	50.995	70.564
	Teacher	82.500	14.262	30	77.175	87.825
	Teammate	72.509	16.397	29	66.271	78.746
R	Friend	78.172	19.990	31	70.840	85.504
	Subordinate	82.636	13.355	30	77.649	87.623
	Teacher	65.217	22.724	28	56.406	74.029
	Teammate	68.598	19.706	29	61.102	76.093
RC	Friend	73.468	21.464	31	65.595	81.341
	Subordinate	77.629	16.178	31	71.695	83.563
	Teacher	85.687	15.474	29	79.801	91.573
	Teammate	74.649	21.244	28	66.411	82.886

TABLE III
DESCRIPTIVES - MDMT: CAPABILITY

Type	Role	Mean	SD	N	95% Credible Interval	
					Lower	Upper
A	Friend	67.670	23.575	27	58.344	76.996
	Subordinate	52.833	25.646	28	42.889	62.778
	Teacher	58.958	27.076	26	48.022	69.895
	Teammate	39.687	24.990	28	29.997	49.378
C	Friend	76.022	20.329	27	67.980	84.064
	Subordinate	57.731	22.962	27	48.648	66.815
	Teacher	77.931	18.980	29	70.712	85.150
	Teammate	68.813	18.803	28	61.522	76.103
R	Friend	73.982	24.245	28	64.581	83.383
	Subordinate	75.567	21.542	30	67.523	83.611
	Teacher	64.935	20.352	28	57.043	72.826
	Teammate	69.375	23.928	28	60.097	78.653
RC	Friend	73.022	23.479	31	64.409	81.634
	Subordinate	73.463	21.374	29	65.332	81.593
	Teacher	83.652	16.137	29	77.514	89.790
	Teammate	73.074	20.313	27	65.039	81.110

TABLE VII
DESCRIPTIVES - UNDERSTANDING CONFIDENCE

Type	Role	Mean	SD	N	95% Credible Interval	
					Lower	Upper
A	Friend	54.429	30.098	28	42.758	66.099
	Subordinate	48.138	28.771	29	37.194	59.082
	Teacher	47.000	37.251	31	33.336	60.664
	Teammate	34.125	32.473	32	22.417	45.833
C	Friend	73.103	25.328	29	63.469	82.738
	Subordinate	53.357	31.645	28	41.086	65.628
	Teacher	78.281	27.568	32	68.342	88.221
	Teammate	68.129	31.937	31	56.414	79.844
R	Friend	74.677	26.530	31	64.946	84.409
	Subordinate	78.531	27.645	32	68.564	88.498
	Teacher	61.607	32.536	28	48.991	74.223
	Teammate	65.207	28.496	29	54.368	76.046
RC	Friend	76.656	25.692	32	67.393	85.919
	Subordinate	70.806	29.290	31	60.063	81.550
	Teacher	76.724	23.298	29	67.862	85.586
	Teammate	70.964	28.288	28	59.996	81.933

TABLE VI
DESCRIPTIVES -PERCEIVED INTELLIGENCE

Type	Role	Mean	SD	N	95% Credible Interval	
					Lower	Upper
A	Friend	68.764	20.825	28	60.689	76.839
	Subordinate	69.441	19.998	29	61.834	77.048
	Teacher	70.316	21.261	31	62.517	78.115
	Teammate	56.350	23.958	32	47.712	64.988
C	Friend	82.297	18.329	29	75.324	89.269
	Subordinate	64.979	23.005	28	56.058	73.899
	Teacher	83.231	16.618	32	77.240	89.223
	Teammate	81.548	15.989	31	75.684	87.413
R	Friend	78.729	21.353	31	70.897	86.561
	Subordinate	77.138	21.905	32	69.240	85.035
	Teacher	77.243	19.781	28	69.573	84.913
	Teammate	79.683	19.846	29	72.134	87.232
RC	Friend	82.869	14.228	32	77.739	87.999
	Subordinate	79.516	17.216	31	73.201	85.831
	Teacher	90.090	10.991	29	85.909	94.270
	Teammate	76.914	19.455	28	69.370	84.458

TABLE V
DESCRIPTIVES - MDMT: SINCERITY

Type	Role	Mean	SD	N	95% Credible Interval	
					Lower	Upper
A	Friend	64.025	30.206	27	52.076	75.974
	Subordinate	64.096	25.923	26	53.625	74.567
	Teacher	69.940	24.480	29	60.628	79.252
	Teammate	74.839	23.325	28	65.795	83.884
C	Friend	82.689	22.685	26	73.527	91.852
	Subordinate	69.417	24.860	26	59.375	79.458
	Teacher	83.169	19.150	31	76.145	90.194
	Teammate	70.399	21.474	29	62.231	78.568
R	Friend	83.325	22.229	30	75.025	91.625
	Subordinate	81.468	21.233	31	73.680	89.256
	Teacher	60.279	29.912	26	48.197	72.361
	Teammate	78.096	23.492	27	68.803	87.389
RC	Friend	80.328	22.405	31	72.110	88.546
	Subordinate	76.524	25.684	31	67.103	85.945
	Teacher	85.554	18.855	26	77.939	93.170
	Teammate	71.558	24.411	26	61.698	81.417