Toward Hybrid Relational-Normative Models of Robot Cognition

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

ABSTRACT

Most previous work on enabling robots' moral competence has used norm-based systems of moral reasoning. However, a number of limitations to norm-based ethical theories have been widely acknowledged. These limitations may be addressed by role-based ethical theories, which have been extensively discussed in the philosophy of technology literature but have received little attention within robotics. My work proposes a hybrid role/norm-based model of robot cognitive processes including moral cognition.

ACM Reference Format:
Ruchen Wen. 2021. Toward Hybrid Relational-Normative Models of Robot Cognition. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3434074.3446353

1 INTRODUCTION AND MOTIVATION

Malle and Scheutz argue that to enable moral competence in social robots, robots need (1) a system of moral norms (moral core); and the ability to use those norms for (2) moral cognition (to generate emotional responses to norm violations and make moral judgements), (3) moral decision making and action (to conform their own actions to the norm system), and (4) moral communication (to generate morally sensitive language and to explain their actions) [17, 18]. This framework is explicitly grounded only in norm-based ethical theories (e.g., deontology). In contrast, we argue for a hybrid framework incorporates both norms and roles into all four of these dimensions of moral competency, as well as into other aspects of robot cognition.

Since most methods for enabling robots moral competence have been based on deontological principles [3] in which the morality of an action depends solely on its consistency with well-specified moral norms [11], it is reasonable that Malle and Scheutz's also ground their work in the concept of norms. However, norm-based ethical theories have philosophical and computational limitations, such as struggling to "accommodate the constant flux, contextual variety, and increasingly opaque horizon of emerging technologies and their applications" [31]. To address these concerns, philosophers of technology have been exploring underrepresented ethical traditions and looking for new perspectives on robotics. Role-based and relational ethical theories, for example, have been discussed in the philosophy of technology literature. For instance, Coeckelbergh discusses the need to focus on moral considerations in human-robot relations rather than on the moral status of humans and robots alone. He also offers an alternative, social-relational approach to moral consideration, which re-frames the issue by shifting the focus from individual ontology to social-relational ontology [7]. Additionally, compared to traditional norm-based approaches, which emphasize epistemological aspects of moral action (e.g., what is good or bad), role-based approaches emphasize ontological aspects of moral learning (e.g., how to become good) [2, 4, 23, 24].

This role-based approach shares some properties with virtue ethics, which has also been discussed in the robot ethics literature [1, 6, 13, 22]. Virtue ethics (e.g., Aristotelianism) primarily focus on the virtues of moral agents themselves [14]. In contrast to virtue ethics, the role-based ethics of Confucianism argues that moral norms and virtues are derived from the social roles humans assume, and social roles in turn are determined by the social relationships humans have with others [21]. Confucian ethics advocate for a relational ontology in which agents never cultivates virtues solely by themselves, and instead becomes virtuous while actively living their social roles through everyday interactions with others [2]. We argue that not only that interactive robots would benefit from a Confucian role ethics approach, but also that interactive robots themselves present a unique opportunity to satisfy classic Confucian ethical goals (e.g., moral education), through the opportunity to encourage the cultivation of others' moral selves [2, 24, 37].

My work proposes a hybrid approach that combines both role ethics and normative ethics to produce a hybrid perspective on every stage of Malle and Scheutz's framework, as well as on other non-moral yet role-based aspects of robot cognition. In the following section, I will discuss this approach in more detail, and summarize my previous, ongoing, and future work on each of the stages.

2 PROPOSED APPROACH AND PROGRESS

2.1 Moral Core

In Malle and Scheutz's framework, the *moral core* is a system of norms and the language and concepts to communicate about these norms [18]. Their work has also discussed the importance of norm acquisition, representation, and contextual activation. One of my current research aims is to develop a set of logical representations that can be used for both role-based and norm-based moral reasoning. While there has been extensive research on norm representations, these works are based on deontology and often use deontic operators to indicate the permissibility of actions [5, 12, 19]. Even

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03.
https://doi.org/10.1145/3434074.3446353

works grounded in virtue ethics [13] have not used representations of roles or relationships.

The role-based approach argues that humans are relational and assume different societal roles [2, 24], and that moral responsibilities are often prescribed by the roles assumes in specific relationships in specific contexts [38]. Thus, from a role ethics theoretic perspective, a robot deciding whether to perform an action must determine whether that action is benevolent with respect to the roles it plays in relation to others, especially those affected by the action in question. For robots to perform this type of reasoning, they require suitable knowledge representations. I am currently developing such representations [36] to enable robots to represent agents, roles, relationships, and actions, and a way of quantifying or otherwise reasoning about the (possibly normative) benevolence of actions with respect to roles and/or relationships.

2.2 Moral Cognition and Moral Decision-Making

After developing a role-oriented moral core, my next step is to enable role-oriented moral cognition and moral decision-making. Moral cognition should allow a robot to reason about whether an action is acceptable. Moral decision making uses this capability when the robot is making its own decisions. Once I have finalized my knowledge representations, I will develop algorithms that enable robots to use this hybrid framework to evaluate the actions performed by others, the actions proposed by others, and the actions the robots themselves are considering.

Most existing computational models for robotic moral decision making are norm-based [8, 16]. From a role ethics perspective, a harmonious society is based on the conscientious fulfillment of the duties demanded by one's assigned roles [10]. Thus, a role-oriented approach to moral cognition and decision making should be sensitive to and enable reflection on the moral responsibilities prescribed by the role(s) a person assumes in a specific context.

We plan to integrate our moral cognition and reasoning module with the Distributed Integrated Affect, Recognition and Cognition (DIARC) architecture [27–29], which currently only considers whether an action is listed as unacceptable and whether states that would be immediately achieved by that action are listed as forbidden.

2.3 Moral Communication

As Malle and Scheutz argue, the cognitive tools that enable moral judgment and moral decision making are important, but they are not sufficient to achieve the socially most important function of morality, which is to regulate the behavior of others [18]. From previous research on Confucian robot ethics [39], we believe that a role-based communication strategy may be particularly effective at inviting human teammates to cultivate self-reflective moral learning, thereby creating not only reliable and efficient human-robot interactions, but also a better moral ecosystem for robots and their human teammates.

To evaluate this hypothesis, we have been conducting human-subjects experiments to investigate the effects of role-based and norm-based moral language in different contexts [15, 32]. In our most recent study, we evaluated the effectiveness of norm-based and role-based moral communication strategies in encouraging compliance with norms grounded in role expectations, in a crowdsourcing context. Our results suggested that: (1) reflective exercises may increase the efficacy of role-based moral language and thus promoting people's task performance; and (2) performing immediate moral practice after receiving role-based moral interventions could help peoples' role-centric moral development by promoting positive attitudes towards behaviors emphasised by the role-grounded moral norms used in such interventions.

In addition, in ongoing work we are collaborating with a United States officer training academy to investigate how differences between these two communication strategies might vary according to contextual factors, due to the differences between military and civilian populations in terms of emphasis on both norms and roles.

2.4 Pragmatic Social Communication

Research has shown that people perceive robots not only as moral agents [9] but also as social actors [20], and thus, we have been extending the role-based approach in our proposed framework to a broader social context. In this case, robots are conforming not only to a system of moral norms, but also a set of sociocultural norms constrained by environmental and social context. We are interested in exploring how robots can learn and use sociocultural linguistic norms, and the relation between these norms and the contextual roles that activate the norms, to understand the intentions of their human teammates and to comply with those learned norms themselves. Specifically, we focus on *Indirect Speech Acts (ISAs)* [30]. People do not always directly express their intentions, especially in contexts that have strong sociocultural norms, conventions, and contracts. In such contexts, humans typically phrase their language as *ISAs*, in which the speech act's literal meaning does not match its intended meaning. For example, in a restaurant, people would typically phrase their request as "Could I have some water" instead of "Get me some water". While this utterance may literally be a request for information, listeners are effortlessly and instinctively able to infer the speaker's true intent, i.e., for the listener to bring them some water, because the use of this conventionally indirect phrasing is itself a sociocultural norm that speakers are expected to follow based on their social role. Accordingly, robots need to be capable of understanding these sociocultural linguistic norms, and the roles that activate them, to appropriately infer the intended meanings behind their teammates' utterances.

Our work [33] addressed this topic by showing how Dempster-Shafer Theoretic norm learning [25, 26] could be used to learn appropriate uncertainty intervals for robots' representations of sociocultural politeness norms surrounding the use of ISAs [34, 35], in a way that is sensitive to the contextual roles that activate those norms. We are currently working to integrate these learned norms within the pragmatic norm base of the DIARC architecture [27–29], and to assess the fluidity and success of robots that use these role-sensitive sociocultural linguistic norms.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Keith Abney. 2012. Robotics, ethical theory, and metaethics: A guide for the perplexed. *Robot ethics: The ethical and social implications of robotics* (2012), 35–52.

[2] R. T. Ames. 2011. *Confucian role ethics: A vocabulary.*

[3] Susan Leigh Anderson and Michael Anderson. 2011. A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care. In *Proc. 12th AAAI Conf. on HRI in Elder Care*. 6. http://dl.acm.org/citation.cfm?id=2908724.2908725

[4] Adam Briggle and Carl Mitcham. 2012. *Ethics and Science: An Introduction.* https://doi.org/10.1017/CBO9781139034111

[5] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. 2006. Toward a General Logicist Methodology for Engineering Ethically Correct Robots. *Intel. Sys.* (2006).

[6] Massimiliano L Cappuccio, Eduardo B Sandoval, Omar Mubin, Mohammad Obaid, and Mari Velonaki. 2020. Can Robots Make us Better Humans? *International Journal of Social Robotics* (2020), 1–16.

[7] Mark Coeckelbergh. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12, 3 (2010), 209–221.

[8] Morteza Dehghani, Emmett Tomai, Kenneth D Forbus, and Matthew Klenk. 2008. An Integrated Reasoning Approach to Moral Decision-Making. In *AAAI*.

[9] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14, 3 (2004), 349–379.

[10] Daniel K. Gardner. 2014. *Confucianism: A Very Short Introduction.* Oxford University Press.

[11] Bertram Gawronski and Jennifer S Beer. 2017. What makes moral dilemma judgments "utilitarian" or "deontological"? *Social Neuroscience* 12, 6 (2017), 626–632.

[12] Aditya Ghose and Tony Bastin Roy Savarimuthu. 2012. Norms as objectives: Revisiting compliance management in multi-agent systems. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. Springer, 105–122.

[13] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Vasanth Sarathy. 2019. Toward the engineering of virtuous machines. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 29–35.

[14] Rosalind Hursthouse. 1999. *On virtue ethics.* OUP Oxford.

[15] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. 2021. Robots as Moral Advisors: The E ects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*.

[16] Vigneshram Krishnamoorthy, Wenhao Luo, Michael Lewis, and Katia Sycara. 2018. 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, 282–287.

[17] Bertram F Malle. 2016. Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots. *Ethics and Info. Tech.* (2016).

[18] Bertram F Malle and Matthias Scheutz. 2014. Moral Competence in Social Robots. In *Symposium on Ethics in Science, Technology and Engineering*. IEEE.

[19] Bertram F Malle, Matthias Scheutz, and Joseph L Austerweil. 2017. Networks of Social and Moral Norms in Human and Robot Agents. In *A World with Robots*.

[20] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[21] A. T. Nuyen. 2007. Confucian ethics as role-based ethics. *International philosophical quarterly* 47 (2007), 315–328.

[22] Anco Peeters and Pim Haselager. 2019. Designing virtuous sex robots. *International Journal of Social Robotics* (2019), 1–12.

[23] H. Rosemont Jr. 2015. *Against Individualism: A Confucian Rethinking of the Foundations of Morality, Politics, Family, and Religion (Philosophy and Cultural Identity)*.

[24] Henry Rosemont Jr and Roger T Ames. 2016. *Confucian role ethics: A moral vision for the 21st century?* Vandenhoeck & Ruprecht.

[25] Vasanth Sarathy, Matthias Scheutz, Yoed N Kenett, Mowafak Allaham, Joseph L Austerweil, and Bertram F Malle. 2017. Mental Representations and Computational Modeling of Context-Specific Human Norm Systems.. In *CogSci*.

[26] Vasanth Sarathy, Matthias Scheutz, and Bertram F Malle. 2017. Learning behavioral norms in uncertain and changing contexts. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 000301–000306.

[27] Paul W Schermerhorn, James F Kramer, Christopher Middendorff, and Matthias Scheutz. 2006. DIARC: A Testbed for Natural Human-Robot Interaction.. In *AAAI*, Vol. 6. 1972–1973.

[28] Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. 2013. Novel mechanisms for natural human-robot interactions in the DIARC architecture. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[29] Matthias Scheutz, Tom Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*. Springer, 165–193.

[30] John R Searle. 1975. Indirect Speech Acts. *Syntax and Semantics* 3 (1975), 59–82.

[31] Shannon Vallor. 2016. *Technology and the virtues: A philosophical guide to a future worth wanting.* Oxford University Press.

[32] Ruchen Wen, Boyoung Kim, Elizabeth Phillips, Qin Zhu, and Tom Williams. 2021. Comparing Strategies for Robot Communication of Role-Grounded Moral Norms. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI-LBR)*.

[33] Ruchen Wen, Mohammed Aun Siddiqui, and Tom Williams. 2020. Dempster-Shafer Theoretic Learning of Indirect Speech Act Comprehension Norms. In *Proce. 34th AAAI Conference on Artificial Intelligence*.

[34] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. 2015. Going Beyond Command-Based Instructions: Extending Robotic Natural Language Interaction Capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[35] Tom Williams, Rafael C Núñez, Gordon Briggs, Matthias Scheutz, Kamal Premaratne, and Manohar N Murthi. 2014. A dempster-shafer theoretic approach to understanding indirect speech acts. In *Ibero-American Conference on Artificial Intelligence*. Springer, 141–153.

[36] Tom Williams, Qin Zhu, Ruchen Wen, and Ewart J de Visser. 2020. The Confucian Matador: Three Defenses Against the Mechanical Bull. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*. 25–33.

[37] Pak-Hang Wong. 2012. Dao, harmony and personhood: Towards a Confucian ethics of technology. *Philosophy & technology* 25, 1 (2012), 67–86.

[38] Qin Zhu. 2018. Engineering ethics education, ethical leadership, and Confucian ethics. *International Journal of Ethics Education* (2018), 1–11.

[39] Qin Zhu, Tom Williams, Blake Jackson, and Ruchen Wen. 2020. Blame-laden moral rebukes and the morally competent robot: A Confucian ethical perspective. *Science and Engineering Ethics* 26, 5 (2020), 2511–2526.