On Perceived Social and Moral Agency in Natural Language Capable Robots

Ryan Blake Jackson
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
rbjackso@mines.edu

Tom Williams
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
twilliams@mines.edu

***Abstract*—Providing robots with social behaviors and natural language capabilities causes laypeople to naturally perceive them as social and moral agents. These perceptions necessitate that robots be sufficiently morally competent to avoid averse effects on the human normative ecosystem and human-robot teaming. Crucially, this required moral competence involves not only moral decision making, but also moral communication. In a communicative medium as sensitive as natural language, the phrasing used to convey a message can be just as important as the message itself.**

***Index Terms*—Natural Language Generation; Robot Ethics; Social Agency; Moral Agency; Perception**

I. Introduction

As the fields of artificial intelligence and robotics continue to advance, we will see robots become more pervasive in a broadening variety of tasks and contexts. Since many of these tasks and contexts will involve interaction with non-roboticists, robot designers are increasingly turning to natural language understanding and generation systems to facilitate easy, fluid communication with nearly all people without requiring burdensome training or specialized hardware [1]–[3].

Despite the advantages of communication via natural language, there are dangers and downsides to equipping robots with this communicative medium, especially since computational dialogue systems are still far from perfect. Natural language is extremely complex, nuanced, and context-dependent. Any given utterance may carry many (context-dependent) implications beyond its literal surface-level meaning [4], [5]. Furthermore, any given message that a robot may want to convey could likely be realized in several different utterances, each with the same primary literal meaning, but with different secondary implications. All of this complexity and nuance puts current dialog systems in real danger of generating utterances with unintentional, potentially misleading, implicatures. As social robots are deployed in increasingly morally consequential contexts, (e.g., eldercare [6], mental health treatment [7], childcare [8], and military operations [9], [10])these accidental implicatures can come with real consequences, and ensuring moral communication and proper communication of moral reasoning becomes nearly as important as ensuring moral actions.

Furthermore, as robots become increasingly capable of general-purpose natural language communication, we argue that laypeople perceive them as increasingly socially competent, and extend social expectations and judgments to robots as perceived social agents and community members. We further argue that the community membership and social status naturally ascribed to social and communicative robots gives rise to a type of moral agency that goes beyond the simple expectation of adhering to moral norms; the robots' social agency grants them a role in the ongoing communal shaping, creating, and enforcing of moral norms, in addition to simply being bound by preexisting community norms. Thus, the interaction of moral and social agencies in robots grants them uniquely powerful normative influence, which necessitates moral competence and communicative sensitivity if that influence is to be wielded responsibly.

Because humans tend to perceive this social and moral agenthood in social robots regardless of the robot's status as truly social, moral, or agentic in any capacity, we adopt the term "perceived moral agency" (PMA) from related work [11], and use a corresponding notion of "perceived social agency" (PSA). We discuss and justify these concepts in more depth in Section II, and differentiate them from asocial moral agency and amoral social agency in Section III. Sections V and IV discuss how the combination of PSA and PMA give robots the power to shape moral norms as community members, and the corresponding necessity for moral and communicative competence. Finally, Section VI concludes this paper with a brief summary and cursory examination of the ontological distinction between human and robot agenthood, despite their shared status as moral and social agents.

II. Perceived Social Agency and Perceived Moral Agency

A large body of research suggests that humans may naturally perceive machines (not just robots) as social actors [12], [13]. This leads people to behave socially towards machines by, for example, applying politeness norms to computers [12].

The research reported in this document was performed in connection with Contract Number W911NF-10-2-0016 with the U.S. Army Research Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of the U.S. Army Research Laboratory, or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Given this human propensity to interact in fundamentally social ways with computers that possess neither feelings nor senses of self nor motivations, it is perhaps unsurprising that social behaviors and social perceptions also manifest in interactions with robots, which are often deliberately designed to be prosocial and anthropomorphised.

Social agenthood appears to be perceived in machines to varying degrees depending on the nature of the machine [13]. For example, language-capable robots are expected to be more socioculturally aware than mute robots, regardless of actual social competence [14]. However, humans tend to project social agency onto, and behave socially towards, even the most basic robots, like Roomba vacuum cleaners [15]. Robots that are more complex, autonomous, anthropomorphic, and communicative can prompt even stronger perceptions of social agency. In fact, recent work has raised concerns that humans may bond so closely with robotic teammates in military contexts that their attachment could jeopardize team performance by causing humans to prioritize the well-being of their robot teammates over mission completion [10]. Furthermore, experiments have shown that humans grant robots in-group social status and apply social categorization processes to robots [16].

A number of researchers have pointed out the dangers of this perceived social agency. Specifically, they claim that human-robot social relationships are likely to be uni-directional; perceived by humans but not truly reciprocated by robots, leading to potential for emotional harm or manipulation at the hand of those robots [15]. Increasing the moral competence of robots would likely alleviate some of these concerns, but, as we will now discuss, PMA without actual moral competence could actually exacerbate these dangers. We now turn our discussion to PMA and its relationship with PSA.

Social robots are also perceived as *moral* agents in that they are expected to behave in accordance with human moral norms. That is to say, people naturally extend moral judgments and blame to robots for their actions similarly to how they would to other people (though perhaps to a lesser extent and according to differing moral philosophies) [17], [18]. Indeed, enabling robots with the requisite moral reasoning capacity to appropriately apply and follow moral norms is an active research area.

In fact, by some definitions of moral agency, robots and other artificial actors can be truly morally agentic, with or without social behavior [19]. By possessing agency, or giving the impression of agency by displaying interactivity, autonomy, and adaptability, robots qualify as sources of moral action (i.e., actions that can be good or bad), and are thus arguably moral agents. However, in this paper, we concern ourselves only with the *perception* of moral agency.

III. Social and Moral Agencies as Independent

We now discuss the extent to which PSA and PMA can manifest in machines independent of one another. We believe that some machines, including robots, are largely perceived as asocial moral agents (PMA without PSA), while others are seen as amoral social agents (PSA without PMA). Although, for the most part, language capable social robots do not fall in either of these groups, we believe that they are worth presenting as points of reference for our discussion of the special moral and social niche occupied by language capable robots.

Some embodied autonomous entities are popularly ascribed some form of PMA without behaving socially or even possessing the capacity for communication outside of a narrow task-based scope. We call this type of PMA "asocial moral agency", and use autonomous motor vehicles as the quintessential example of asocial perceived moral agents.

Autonomous motor vehicles are clearly expected to conform to the legal rules of the road, but they are also expected to engage in extralegal moral decision making and moral reasoning. Furthermore, they are subject to moral judgments and blame for any behavior perceived to violate standing moral norms. There are myriad articles, both in popular culture and in academia, contemplating whether and how autonomous cars should make decisions based on moral principles (e.g., [20]). Questions like "in an accident, should the car hit a school bus to save its own passenger's life? Or should it hit the barrier and kill its passenger to save the school children?" have taken hold of people's imaginations and proliferated wildly. Regardless of the actual usefulness of such questions, it is clear that autonomous cars are being ascribed not only moral agency, but also moral responsibility.

We can also consider that autonomous vehicles, once they become more ubiquitous, might change the norms governing human driving behavior. For example, if all autonomous vehicles on a road adopt a uniform following distance, this behavior might influence human drivers sharing the road to do the same, or to alter their driving behavior in some other way.

However, the potential normative influence of autonomous cars is distinct from that of social robots in that it is passive, incidental, and unintentional. In contrast, as discussed below, social robots can exert their normative influence purposefully and actively. Therefore, despite previous research showing no difference in magnitude of PMA between various *social* machines (including social robots) [11], we posit that these social machines are more morally agentic than similarly intelligent asocial machines like autonomous cars.

We can also consider cases where, depending on behavior, robots could be perceived as amoral social agents. Social robots that do not have the ability to act on their environment in any meaningful capacity may be physically unable (or barely able) to produce moral action. Such a robot could, however, be the recipient of moral action (a moral patient). Especially given the inverse relationship between moral agency and moral patiency [21], this robot would be considered minimally morally agentic. As an example, consider MIT's Kismet robot, which is expressive, (non-linguistically) communicative, and social, but largely helpless and incapable of acting in an extra-communicative capacity.

IV. Social and Moral Agencies Interacting: The Need for Moral Communication

The PMA of social robots interacts with their PSA to create implications beyond robots simply being held to communal standards of moral behavior. Because of their PSA, potential in-group social status, and perceived community membership, social robots can (and do) actively participate in creating, shaping, and enforcing the norms that inform human morality.

Empirical studies in behavioral ethics have shown that human morality is dynamic and malleable [22], and a society's moral norms are defined and developed both by human community members and the technologies with which they interact [23]. Social robots occupy an interesting sociotechnical niche at the intersection between agentic community member and technological tool. This position enables them to wield a more significant normative influence than many other technologies. For example, robots have been shown to hold measurable persuasive capacity over humans, both via direct persuasion and implicit pressure towards behavioral conformity [24], [25].

We therefore believe that language capable robots are unique among technologies in their ability to take an active, purposeful, and autonomous role in shaping human moral norms (or human application of moral norms). However, this capability is a double-edged sword. On the one hand, robots of the future could productively influence the human moral ecosystem by reinforcing desirable[1] norms and dissuading norm violations. On the other hand, today's imperfect natural language dialogue systems open the door for robots to inadvertently and detrimentally impact the human moral ecosystem through miscommunications and unintended implicatures. Much of our recent work concerns this issue.

Given the behavioral expectations and normative influence imparted to social robots by their combined PSA and PMA, it is crucial that their natural language software ensures moral communication and proper communication of moral reasoning, especially in serious or morally consequential contexts. The power to transfer or alter norms comes with the responsibility to do so in a morally sensitive manner. However, as discussed previously, the intricacies of natural language and the breadth of contexts in which robots will interact with people make this a difficult goal. We will now discuss some shortcomings of current dialogue systems, and explore the potential benefits that morally sensitive communication might bring in the future.

V. Our Recent Empirical Results

One aspect of current language-capable robot architectures that risks miscommunication is clarification request generation. For performance reasons, current language pipelines generate clarification requests as soon as ambiguity is identified in a human request. This reflexive action preempts any moral reasoning modules that the robot might have until the ambiguity is resolved. In most situations, this method is not problematic. However, if a human request is both ambiguous and morally objectionable, the robot risks accidentally implying a willingness to violate a moral norm by requesting clarification.

As an example, consider the command "break her phone." Barring extenuating circumstances, compliance with this command is presumably morally impermissible. However, if the noun phrase "her phone" is an ambiguous reference (e.g., the robot knows about two conversationally salient females with phones), then the robot would seek clarification with an utterance along the lines of "Should I break Shauna's phone or Leah's phone?", and, in so doing, imply a willingness to break at least one of the listed phones. The robot would generate this utterance even if moral reasoning modules would prevent it from ever actually breaking a phone because, when it requests clarification, the robot is not actually considering any action; it is simply trying to resolve ambiguity to proceed with sentence processing.

In a series of recent studies [26], [27], we showed written or filmed human-robot clarification dialogues to human subjects. The control group was shown a clarification dialogue regarding a morally benign request, while the experimental group was shown a structurally identical dialogue about a morally objectionable request. Before and after exposure to the assigned clarification dialogue, participants reported both the degree to which they found the morally objectionable action permissible and their impression of the degree to which the robot thought the morally objectionable action was permissible. Our results indicated that generating clarification per the status quo did cause robots to miscommunicate their intentions by erroneously implying willingness to violate the relevant norm. Furthermore, and perhaps more worryingly, the clarification dialogue weakened humans own perceptions of the strength of that norm, at least within the examined experimental context. In other words, participants reported higher levels of perceived permissibility for the morally problematic action after reading the clarification dialogue pertaining to that action. We found evidence that these effects occur for both human interactants and third-party observers, and that these results are at least somewhat robust to social distance.

These recent studies show how adversarial contexts or inputs can cause dialogue systems to generate unintended implications, and how these implications can mislead humans, damage their perceptions of the robot, and even alter human moral reasoning. We believe that the powerful consequences of the unintended implications are products of the robot's perceived moral and social agency.

In another recent study, we examine the importance of phrasing in morally charged speech acts [28]. Specifically, we look at verbal robotic noncompliance. Robots should not blindly follow every command that they receive. If a human requests something that the robot knows to be immoral, then the robot should refuse to comply. However, as a perceived moral and social agent, the phrasing that the robot uses in its refusal is important to its position within

[1]Given humanity's lack of consensus within moral philosophy, whether any given norm is desirable may be a matter of opinion. We leave this matter to moral philosophers, but opine that equipping a robot with any single framework of morality should be done only after considerable deliberation.

the community and the continued efficacy and amicability of human-robot teaming. Our experiment pairs two human requests carrying different levels of moral permissibility with two robotic refusals carrying different levels of politeness (or face threat, see [29]). We found evidence that the degree of politeness theoretic face threat in a command rejection should be proportional to the severity of the norm violation motivating that rejection. Specifically, we saw significant decreases in the robot's likeability when the less polite refusal was paired with the less morally objectionable command. Subjects also rated the robot as too harsh in this pairing, and, critically, as not harsh enough when the more polite refusal was paired with the more morally objectionable command.

Overall, these studies points to a gap between the current capabilities of language generation systems to generate morally sensitive language, and the responsibility to communicate morally that robots, as perceived moral and social agents, should uphold given their profound normative influence.

VI. Concluding Remarks

Providing robots with social behaviors and natural language capabilities induces perceptions of moral and social agency which necessitates that we also provide them with moral competence to avoid averse effects on the human normative ecosystem and human-robot teaming. Crucially, this required moral competence involves not only moral decision making, but also moral communication. In a communicative medium as sensitive as natural language, the phrasing used to convey a message can be just as important as the message itself.

Despite people's shared status as moral and social agents, we cannot necessarily prescribe human communicative strategies and behaviors to robots. Previous research indicates that robots, and other social machines, are perceived as ontologically distinct from humans in terms of moral agency [11]. Therefore, what is best for a human to say or do in any given situation might not be the same for a robot. Further research is required to discover the implications of (perceived) moral and social agency specific to robots, the norms that apply to robots, the norms that humans will apply to themselves in interacting with robots, and how robots will fit into our communities as norm shapers.

Though we have evidence for an ontological distinction between humans and robots as moral and social agents, it is not yet clear exactly where the differences (and similarities) will manifest. It is common for human-robot interaction researchers to run experiments and report results without including any human-human interaction point of reference. In our clarification request generation studies described previously, for example, we did not collect any data to discover whether the robot's normative influence would be less or more powerful than that of a human displaying the same behavior. We will require such points of reference if we are to fully understand how the emerging PMA and PSA of robots relate to moral and social agency in humans, and how humans and robots ought to socialize and communicate with each other.

References

[1] M. Scheutz, B. Malle, and G. Briggs, "Towards morally sensitive action selection for autonomous social robots," in *Proc. of RO-MAN*, 2015.

[2] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, 2015.

[3] C. Matuszek, "Grounded language learning: Where robotics and nlp meet." in *Proceedings of IJCAI*, 2018.

[4] P. Grice, "Logic and conversation," in *Syntax and Semantics*, 1975.

[5] S. C. Levinson, *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.

[6] K. Wada and T. Shibata, "Living with seal robots – its sociopsychological and physiological influences on the elderly at a care house," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 972–980, 2007.

[7] B. Scassellati, H. Admoni, and M. Mataric, "Robots for use in autism research," *Annual Review of Biomedical Engineering*, vol. 14, 2012.

[8] N. Sharkey and A. Sharkey, "The crying shame of robot nannies: an ethical appraisal," *Interaction Studies*, vol. 11, no. 2, pp. 161–190, 2010.

[9] R. C. Arkin, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture," in *Proceedings of HRI*, 2008.

[10] J. Wen, A. Stewart, M. Billinghurst, A. Dey, C. Tossell, and V. Finomore, "He who hesitates is lost (...in thoughts over a robot)," in *Proceedings of the Technology, Mind, and Society*. ACM, 2018.

[11] J. Banks, "A perceived moral agency scale: Development and validation of a metric for humans and social machines," *Computers in Human Behavior*, vol. 90, pp. 363 – 371, 2019.

[12] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1994, pp. 72–78.

[13] I. Straub, "'it looks like a human!' the interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription," *AI & SOCIETY*, vol. 31, no. 4, pp. 553–571, 2016.

[14] R. Simmons, M. Makatchev, R. Kirby, M. K. Lee *et al.*, "Believable robot characters," *AI Magazine*, no. 4, 2011.

[15] M. Scheutz, "13 the inherent dangers of unidirectional emotional bonds between humans and social robots," in *Robot Ethics*. MIT Press, 2011.

[16] F. Eyssel and D. Kuchenbrandt, "Social categorization of social robots: Anthropomorphism as a function of robot group membership," *British Journal of Social Psychology*, no. 4, 2012.

[17] P. H. Kahn, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. Gary, A. L. Reichert, N. G. Freier, and R. L. Severson, "Do people hold a humanoid robot morally accountable for the harm it causes?" in *Proceedings of HRI*, 2012, pp. 33–40.

[18] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice one for the good of many?: People apply different moral norms to human and robot agents," in *Proceedings of HRI*, 2015.

[19] L. Floridi and J. Sanders, "On the morality of artificial agents," *Minds and Machines*, vol. 14, no. 3, pp. 349–379, 2004.

[20] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," *Science*, vol. 352, no. 6293, pp. 1573–1576, 2016.

[21] K. Gray and D. M. Wegner, "Moral typecasting: Divergent perceptions of moral agents and moral patients," *Journal of Personality and Social Psychology*, vol. 96, no. 3, 2009.

[22] F. Gino, "Understanding ordinary unethical behavior: Why people who value morality act immorally," *Current opinion in behavioral sciences*, vol. 3, pp. 107–111, 2015.

[23] P.-P. Verbeek, *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press, 2011.

[24] G. Briggs and M. Scheutz, "How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress," *International Journal of Social Robotics*, 2014.

[25] J. Kennedy, P. Baxter, and T. Belpaeme, "Children comply with a robot's indirect requests," in *Proceedings of HRI*. ACM, 2014, pp. 198–199.

[26] R. B. Jackson and T. Williams, "Robot: Asker of questions and changer of norms?" in *Proceedings of ICRES*, 2018.

[27] ——, "Language-capable robots may inadvertently weaken human moral norms," in *Proceedings of alt.HRI*, 2019.

[28] R. B. Jackson, R. Wen, and T. Williams, "Tact in noncompliance: The need for pragmatically apt responses to unethical commands," in *Proceedings of AIES*, 2019.

[29] P. Brown and S. Levinson, *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.