Computer Ethics - Philosophical Enquiry (CEPE) Proceedings

Volume 2019 *CEPE 2019: Risk & Cybersecurity* Article 18

5-29-2019

The Incorporation of Moral-Development Language for Machine-Learning Companion Robots

Patrick Lee Plaisance
Pennsylvania State University

Joe Cruz
Pennsylvania State University

Follow this and additional works at: https://digitalcommons.odu.edu/cepe_proceedings

Part of the Applied Ethics Commons, Communication Technology and New Media Commons, and the Ethics and Political Philosophy Commons

Custom Citation

Plaisance, P. L., & Cruz, J. (2019). The incorporation of moral-development language for machine-learning companion robots. In D. Wittkower (Ed.), *2019 Computer Ethics - Philosophical Enquiry (CEPE) Proceedings*, (15 pp.). doi: 10.25884/kxmq-cd24 Retrieved from https://digitalcommons.odu.edu/cepe_proceedings/vol2019/iss1/18

This Paper is brought to you for free and open access by ODU Digital Commons. It has been accepted for inclusion in Computer Ethics - Philosophical Enquiry (CEPE) Proceedings by an authorized editor of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

The incorporation of moral-development language for machine-learning companion robots

Patrick Lee Plaisance
Pennsylvania State University

Joe Cruz
Pennsylvania State University

Abstract

Among the ongoing debates over ethical implications of artificial-intelligence development and applications, AI morality, and the nature of autonomous agency for robots, how to think about the moral assumptions implicit in machine-learning capacities for so-called companion robots is arguably an urgent one. This project links the development of machine-learning algorithmic design with moral-development theory language. It argues that robotic algorithmic responses should incorporate language linked to higher-order moral reasoning, reflecting notions of universal respect, community obligation and justice to encourage similar deliberation among human subjects.

Keywords: AI, robots, robot ethics, AI ethics

Robots and intelligent machines are increasingly assisting and often displacing human workers in various sectors including manufacturing, finance, education, and health care. At first, the robots' capabilities were limited to helping expedite and standardize particular processes, such as enhancing assembly-line processes or augmenting open-heart surgery. In most contexts, machines were interacting with other machines, and human interventions were limited to installing, operating, and fixing them. Advances in artificial intelligence, however, have resulted in the proliferation of robots into nearly all facets of human private and social life. Human-machine interactions increasingly mimic human-human communication, and in some cases, aim to substitute human socialization altogether. Some of these changes occur in the health-care industry, where human-machine interactions could produce therapeutic and even life-saving results.

The development of intelligent mechanical devices for companionship and care is driven in large part by rapidly changing demographics. First, the general population is becoming older and living longer (Older People Projected to Outnumber Children, 2018). Second, as people age, they are more likely to live alone (West et al., 2014). Third, older individuals tend to experience mental-health problems at higher rates than the general population. In fact, men aged 85 and older are four times more likely to commit suicide than younger men (Centers for Disease Control and Prevention, 2008). Lastly, perhaps as a byproduct of the previous factors, the health-care industry's need for caregivers is expected to grow exponentially in the following decade (Home Health

Aides and Personal Care Aides, n.d.). Aiming to capitalize on this opportunity, companies have produced intelligent devices, or "robot companions," capable of performing many caregiving and therapeutic tasks. They often provide supervision and a sense of companionship to people with particular health and social needs, such as children and the elderly.

The general public reports mixed feelings about introducing robot companions into their daily routines. Approximately 40 percent of Americans say they would be interested in purchasing a robot caregiver, while close to 60 percent express more skepticism about the idea (Smith & Anderson, 2017). Yet despite the public's cautious interest, the robot companion market is projected to become a visible staple of the health-care industry. Orders for robot companions increasingly surpass those used for logistics and manufacturing (Business Insider, 2015). In fact, the market is expected to be worth more than $1.5 billion by 2020, and nearly $2 billion a decade later (Value of social and entertainment robot market, n.d.). Because providing certain kinds of health-care has traditionally depended on human-to-human interaction, the increasing use of robots could reshape communication dynamics and socialization.

While research concerning robot-human interactions spans across decades and fields, recent technological advances and the relatively low cost of such devices signal a shifting landscape. The development and deployment of robot companions raise serious moral questions that have not been adequately explored. What do we mean by morality when it comes to machine-learning algorithms? How are issues of value, virtue and moral principles factored into machine-learning algorithms that will be "talking" to humans with real needs? What kind of "conversation" will be generated between machine-learning androids and humans struggling with loneliness, depression and perhaps even tendencies of self-harm or violence toward others? Companion robots are being commissioned to replace—not assist—human caregivers, effectively rupturing socialization stages for three vulnerable populations: children, the elderly, and the mentally disabled. The emergence of this "artificial socialization" raises questions concerning how robots affect people's privacy, ability to socialize, mental and physical health, and agency. This phenomenon requires us to consider the moral framework that guides what robots are programmed to say to help improve their custodees' well-being. This project aims to propose the incorporation of moral reasoning language and theory in the development of machine-learning algorithms, to ensure the likelihood that android "talk" emphasizes higher-order moral thinking rather than merely reinforce other anti-social, self-serving or self-defeating impulses that robots may "hear" from their human "patients."

Moral-development theory tells us that, through childhood and into emerging adulthood (and often beyond), individuals engage in morally immature thinking that emphasizes egocentrism, high levels of individualism, and relativistic thinking. As more morally developed adults, most individuals tend to grow out of these features and recognize value of reciprocity and pro-social behavior. Finally, higher moral thinking emphasizes cooperation, social duties, the common good, and the internalization of moral principles such as universal respect and justice. Specific language, or moral narratives, reflect each of these levels of moral development (Kohlberg, 1971; Rest et al., 2000). This project articulates why machine-learning algorithm developers must take the language of moral-development theory into account, and it proposes some concretes examples that can “guide” robot-human conversations to emphasize higher-order moral thinking when it would be advantageous to human subjects to do so.

The uncanny valley

Concerns about the rapid integration of artificial beings into social milieus typically address aesthetics and, specifically, a robot’s physical resemblance to humans. Mori (2012) hypothesizes that, when humans encounter objects that harbor anthropological characteristics, they enter an eerie middle-ground called the “uncanny valley,” where human appearance and mechanical behavior intersect. In Mori’s “valley,” a humanoid object would trigger emotions of unease in human observers. Hence, some robot designers argue that pushing robots out of this valley with models that erase the lines between human and machine should be a priority. Hiroshi Ishiguro’s work aims to achieve this goal and to make robot-human interactions feel as natural as possible for people.

As the director of Intelligent Robotics Laboratory at Osaka University, Ishiguro oversees artificial agent designs and programming. Ishiguro’s androids have attracted significant attention because of their unique resemblance to humans, including their precise replication of body and facial movements. His laboratory refers to these designs as “actroids,” androids with visually recognizable human likenesses. Most versions typically have been modeled after an average young Japanese woman, and they have been programmed to mimic lifelike functions such as blinking, speaking and breathing. Some models even use an AI that allows them to respond to physical contact. Ishiguro is perhaps notoriously known for designing and carrying around the Geminoid H1, an android that looks like him. Like Ishiguro, the Geminoid H1 has a thin build, thick eyebrows, piercing black eyes, real black hair plucked from Ishiguro’s scalp, and is dressed in a black long-sleeve shirt and black jeans. It mimics breathing, moves his jaw and lips to simulate speech, and nods his head, even though it can only perform these functions while sitting down. It is operated remotely, and sometimes Ishiguro uses a microphone to project his voice unto the Geminoid’s facial movements. The term “geminoid,” which Ishiguro coined, stems from *geminus*, the Latin word for twin.

Ishiguro indicates that his inventions, particularly the Geminoid H1, help researchers understand face-to-face interactions between individuals and artificial agents (MacDorman & Ishiguro, 2006). He has stated that he is “genuinely interested in knowing how human beings react to being in the presence of a robot” (Guizzo, 2010). Ultimately, Ishiguro wishes to design robots that move more naturally—in a less “mechanized” way—so that they transcend their uncanniness. When this is accomplished, he argues, human beings will be more likely to recognize robots as human in certain social contexts (Ishiguro, 2006). For now, companion robots cannot convincingly express a range of emotions or move in a non-static way, but they are slowly displacing human beings from contexts in which socialization is an important aspect (Simon, 2017). Machines, Ishiguro stipulates, have historically improved and replaced human abilities, and he is just trying to contribute to this paradigm. Despite how human-like Ishiguro’s androids are, they remain trapped in the uncanny valley. Ishiguro’s main concerns about his androids and the future of robotics are not necessarily motivated by the ethical implications of their social impact. Rather, he focuses on how robot aesthetics elicit various psychological reactions from humans and how they can be improved to look (and act) as human as possible.

Moral development theory

Jean Piaget's (1932) studies about how children approached moral dilemmas serve as a foundation for the field of moral-development research. His conclusions furthered the idea that the development of human moral reasoning is contingent on both cognitive changes and external stimuli. Piaget also conceptualized moral development within a stage-based model, where the early stages consist of rudimentary ideas of moral reasoning and the later stages show more advanced moral judgement. Essentially, in Piaget's view, pre-established cognitions complement and accommodate new cognitions. For example, when children apply moral reasoning to differentiate between right and wrong, they tend to do so from a lens of self-interest. They recognize that rules emerge from a higher (adult) authority and obey them because they fear the severity of a hypothetical punishment. When adolescents encounter situations where a distinction between right and wrong must be drawn, a consideration of norms and rules established by their immediate groups (e.g., family) occurs. They begin to recognize that rules may exist to maintain a social order that benefits them. Adults also employ moral reasoning contingent on group norms, and some of them exhibit higher levels of moral thinking, where rules are assessed and sometimes changed based on their social relevance and adherence to universal moral principles.

Moral development theory chronicles and examines the evolution, as it is shaped by cognitive evolution and environmental circumstances, of moral deliberation. Kohlberg (1971) built upon Piaget's findings and expanded his propositions of a stage-based moral development process by theorizing that individuals transition from self-interest-informed moral reasoning to a more socially-conscientious mindset. Kohlberg's theorizing has become a predominant framework through which moral reasoning is explained and assessed. A moral psychology survey instrument, the Defining Issues Test, is tied to Kohlberg's theory of six stages of moral development; it has been refined and used with hundreds of thousands of subjects over the last four decades to assess the moral reasoning levels of various populations. In Kohlberg's view, individuals develop an increasingly sophisticated moral compass that conceptualizes justice as an ever-expanding scope of concern for others. The six stages in Kohlberg's moral reasoning are divided as follows:

I. Preconventional level ("pre-moral").
At this level, children are aware of cultural norms and interpretations of good and bad, but they base their perception of these elements on their self-interest and the consequences of actions. A philosophy of quid pro quo and quasi hedonism guides their instincts. Additionally, they tend to decide how to act based on the physical power of those who create rules.

Stage 1: *Punishment and Obedience.* Young individuals determine what is good or bad based on the physical consequences of a particular act. They act in ways to avoid punishment.

Stage 2: *Instrumental relativist orientation.* Interactions with other individuals are assessed within a framework of reciprocity. Actions must satisfy one's needs (and, sporadically, the needs of others) and desires.

II. Conventional level.

This level finds immediate groups to be of great importance to individual actions. Individuals develop a sense of kinship to their groups, and thus, respect for social order, its maintenance and justification. Typically, this is the highest level of moral reasoning that most adults achieve.

Stage 3: *Interpersonal accordance.* Interpretations of right and wrong are contingent to a group's consensus of what is "bad" and "good." Intention, not necessarily consequences, gain currency at this stage.

Stage 4: *Law and order*. Individuals feel encouraged to maintain social order by following rules set by various groups. Respect for authority and social institutions is predominant.

III. Postconventional level.

Only a limited number of individuals exhibit this type of moral reasoning. People who evaluate situations at this stage tend to reevaluate rules and the authority that establishes them based on universal moral principles and their value to the common good, independent from one's immediate groups.

Stage 5: *Social contract (within a quasi-utilitarian framework).* The possibility of changing rules, laws and rights based on their "social utility" and personal "values" emerges. Justice inhabits a gray area because individual rights and personal freedom are celebrated, as long as they do not infringe on others' rights.

Stage 6: *Universal ethical principles.* All human beings are elevated as individual persons with inherent dignity, who are equal and deserve respect. Ethical principles are abstract, consistent and harbor a sense of universality.

Moral psychology researchers have suggested that a number of factors influence moral development. The DIT is devoid of gender bias, and level of education, not age, is a primary predictor of moral reasoning. On the other hand, religion (Rest et al., 2000; Parker, 1990) and political ideology, particularly of individuals with relatively conservative views (Rest et al., 2000), tend to predict a low DIT score.

Moral reasoning is contingent on a variety of factors that human beings are exposed to or experience throughout their lives. These include cognitive development, group membership and kinship, formation and maintenance of relationships, education, political leanings, history of civic participation, religion, and life experience. Theorists say moral development operates more like an ecosystem, in which various elements interact and change, and less like a step-by-step process. We make decisions considering a plethora of lived experiences that manifest in various levels of consciousness. With this in mind, it is important to interrogate how a sense of morality can be embedded into artificial intelligence.

Moral artificial agents for companionship

How do we instill moral responsibility into machines that lack the same cognitive development of human beings? Bertolini and Aiello (2018) argue that, first, they should be conceptualized as products, *not* subjects. Robot companions are designed by companies focused on finding ways to improve their products and generate profits. Considering how many companies operate within a surveillance capitalism framework, companion robots could also function as data-gathering devices to both customize subject care and feed third party data banks. Second, AI governs and, to some extent, limits robot companions' care capabilities (Stahl & Coeckelbergh, 2016). Although a robot companion programed to assist a diabetic might be able to remind the patient about taking insulin, it would not be equipped to show empathy for him. Third, robot companions need some kind of moral reasoning language embedded in their programing (Stahl & Coeckelbergh, 2016). However, this issue raises questions about how to design morality for a quasi-autonomous agent (UNESCO, 2017). After all, even though robots are created with a particular AI that guides their actions, in the end, these actions are taken independently from others, based on their previously uploaded programing. Lastly, if robot companions require the ability to make decisions as moral agents, how do we decide which ethical theory informs their actions? Engineers, policymakers, and ethicists will probably disagree on which approach to take and which ethical language to embed on an artificial agent (Bogosian, 2017).

Although the emerging robot companion market has attracted criticism from ethicists, much focus has gone to matters of agency, selfhood, deontology, and ontology, and less to the actual communicative process with humans and its moral dimensions. Because some of these mechanical artifacts care for individuals with serious health risks, it is important to examine the actual linguistic choices used to provide services. But the responsibility of communicating health-improving messages poses numerous ethical challenges for robot companions, their creators, and patients. Some robot companions are classified as self-learning machines, which means they are programmed to learn from repeated exposure to certain experiences. This modality allows them to adapt more easily to changing environments, particularly if the custodee suffers from a complex health condition that requires routine data acquisition or monitoring. For example, patients with PTSD may require a mechanical companion that engages in numerous tasks such as tracking physical movements, contacting loved ones or emergency contacts, crafting a conversation to assess moods, and reminding them to take medicine. Artificial agents created to perform health-related tasks—or at least those currently available in the consumer market—depend on some level of machine-human/human-machine interaction to operate adequately. Humans give commands or signals to machines so that they recognize a particular need. In turn, machines acquire information from humans to customize a pre-programmed service. In other words, human-machine engagement increasingly mimics human-human communication insofar as two or more agents exchange information.

While there are numerous health benefits to his capability, it is important to emphasize the moral challenges that artificial agents pose. Vallor and Bekey (2017) argue that creating robot companions capable of producing and maintaining communication with humans present challenges and opportunities to improve such interactions. First, regardless of the accuracy and efficiency of their algorithms, machines that intervene in human activities warrant a consistent level of human supervision. Both the algorithm-development and human-oversight processes should be at least partially available for consumers to access and understand. Of course, there are legal limitations to this proposition, because in some cases, both aspects of the mechanical architecture of a particular robot may be proprietary. Second, algorithms have shown time and again that they harbor human biases that could disadvantage certain groups. Credit applications used by financial institutions sometimes consider geographic and demographic factors that could exclude people based on ethnicity. For example, robot companions operate using algorithmic languages created by people that could embed biases against the same populations the robots are designed to assist.

Moral language

Clearly, technology has changed the way humans perceive and employ language, and how they use language to refer to or address technology. When humans interact with machines, they engage in what Coeckelbergh (2017) refers to as "language games"—that is, associating technology with other elements of our social milieu and treating it as such. For example, the zoomorphic characteristics of some robot companions, such as PARO and "Joy for All," which resemble a seal and a dog or cat respectively, may endear some humans to speak to them as pets (e.g., "who's a good boy?"). It is now part of the collective lingo to tell someone to "Google" something when they inquire about a particular topic. The phrase "Hey, Siri," commands the iPhone to perform a task. Certain virtual health assistants are programmed to recognize words that may require professional intervention. In other words, technology shapes the same language that mediates our interactions (Coeckelburgh, p. 45, 2017).

How could we expect robot companions to communicate and execute an adequate notion of morality when they are "mindless" agents? La Bossiere (2017) argues that to ascribe moral agency to robots, first we must determine if they harbor a moral status. Moral status consists, according to him, of an agent's ability to reason and express feelings. Reason, applied to an artificial agent, would encompass communicating in a language similar to that of humans. This broader question of moral agency is no doubt critical in several areas of AI development. When might it be valuable or necessary for the development of a moral "profile" or simulation of moral motivation in human-computer interaction? How exactly might these be developed? However, this project is not focused on such questions of moral agency, but instead on how companion robots might cultivate morally-relevant interactions through machine learning, and how we might guide those interactions to reinforce higher-order moral thinking when appropriate. As Paula Boddington cautions:

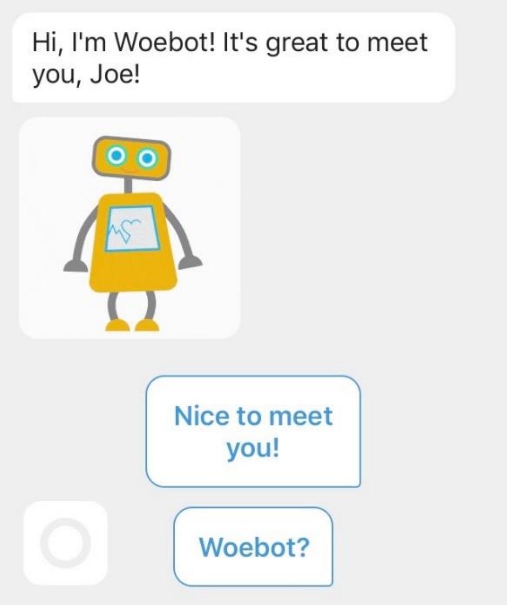
> It's important to remember that AI can take many forms, and be applied in many different ways, so none of this is to argue that using AI will be 'good' or 'bad'. In some cases, AI might nudge us to improve our approach. But

> in others, it could reduce or atrophy our approach to important issues. It might even skew how we think about values (2019).

Boddington points to the ongoing debate over AI applications in medicine. The potential benefits in the development of AI as some sort of “repository for the collective medical mind” are enormous: more efficient diagnoses, systematic elimination of dissenting opinions. However, theorists caution that allowing this could come at great costs. In our tendency to perceive such a tool as infallible, we might well start to foreclose independent thought and clinical experience. Moreover, such machine learning could even be manipulated to aim for treatment or profit targets that benefit special interests rather than patients (Char et al., 2018). Similarly, machine-learning responses to morally significant issues or queries by companion robots poses important questions about how they might “help” people in a state of distress, despair or contemplating harming themselves or others.

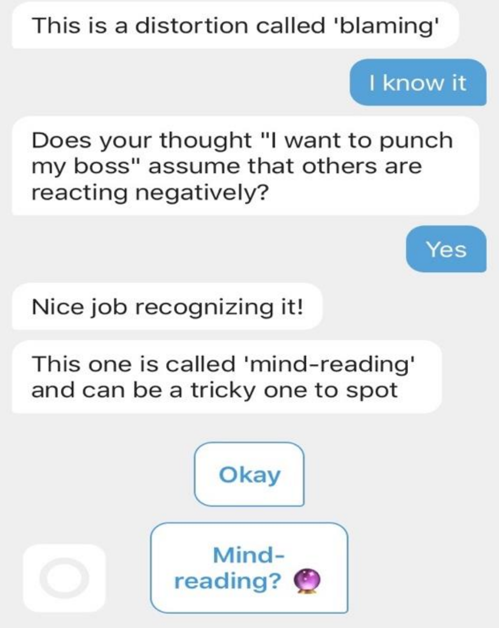
Artificial moral language applications

Although the robot companion market is projected to grow exponentially in the following decades, making some of these devices accessible for many consumers, so far, they remain out of reach for the average person. In the meantime, certain apps for iOS play the role of either companion or quasi-therapist. Woebot is one of these apps (See Figure 1). Its algorithm was fed with cognitive behavioral therapy (CBT) language to help people cope with mental health issues. Although not intended to substitute a real therapist, the AI mimics what an exchange with a mental-health professional would look like.



Hi, I'm Woebot! It's great to meet you, Joe!

Figure 1. Woebot greets user.
Source: Woebot App



This is a distortion called 'blaming'

I know it

Does your thought "I want to punch my boss" assume that others are reacting negatively?

Yes

Nice job recognizing it!

This one is called 'mind-reading' and can be a tricky one to spot

Figure 2. Woebot’s response to a user wanting to “punch their boss”
Source: Woebot App

When informed that the user feels like “punching his boss,” the app reacts by referencing the “mind reading” aspect of CBT, which refers to our tendency to assume that we know what others are thinking (See Figure 2). We do not wish to engage with the psychological value of Woebot’s recommendation. Rather, we suggest that incorporating moral language could analogously help the user’s deliberations on moral issues.

Moral language for AI

Like the Woebot app, whose machine-learning responses are largely based on queries designed to encourage more talk and open up spaces for extended deliberation, generation of responses informed by moral development theory also would be query-driven rather than declarative. This reflects the language of the moral reasoning assessment instrument, the Defining Issue Test[1], which has been used for decades. The stage theory of moral development by Kohlberg, detailed earlier, provides the theoretical basis for the DIT. James Rest, who developed the DIT, reconceptualized Kohlberg’s idea of hard-and-fast stages using schema theory. Rest and colleagues (2000) sought to use the term “schemas” instead of “stages” because they view moral reasoning as a web of interconnected concepts that can be evoked for different situations, rather than a staircase-like progression. Schemas, which are mental shortcuts we routinely use to make sense of information and ideas, are based on our

[1] The Defining Issues Test is copyrighted and requires permission for use. Those interested in details are directed to contact the Center for the Study of Ethical Development at the University of Alabama (https://ethicaldevelopment.ua.edu/).

conclusions we've made about a relevant idea or issue (Fiske & Taylor, 1984). Similarly, we have schemas for types of ethical problems, which we draw upon when confronted with new dilemmas (Rest et al., 2000). Rest and colleagues theorized that if a person has acquired a schema linked to the highest stages of ethical reasoning, statements reflecting that stage on the DIT will activate those schemas; otherwise lower-stage schemas will be drawn upon. Research over the years has suggested that people can generally be identified with one stage in their moral reasoning, but there is fluidity: people draw from lower or higher stages as well depending on the situation they confront. Consequently, rather than a staircase with discrete steps, moral development is best understood as a sliding scale.

The DIT presents subjects with a series of moral dilemmas; for each dilemma, subjects are asked to consider twelve statements or queries that indicate various ways of responding to or making judgments about how to resolve the dilemma. Subjects are asked first to *rate* their perceived importance of each item on a five-point scale to their thinking about the dilemma. They are then asked to identify and *rank* what they perceive to be the four most important items. For each of the dilemmas, there are statements or queries that 1) reflect pre-conventional or conventional moral thinking, 2) post-conventional or higher-order moral thinking, and 3) non-sensical yet plausible moral statements about the dilemma. Researchers can calculate these responses to weed out questionable responses (i.e., those who highly rate/rank the non-sensical statements) and to assess a *P* score, which indicates the percentage of responses by subjects that reflect higher-order moral thinking. The DIT is a robust instrument that has been used with various populations and hundreds of thousands of subjects over the last three decades. Its validity has been assessed for multiple demographic and moral-comprehension criteria cited in more than 400 published articles. The instrument is equally valid for males and females, and meta-analyses show that 30 percent to 50 percent of the variance of DIT scores is attributable to level of education in samples ranging from junior high school students to Ph.D. holders. Cronbach's alphas for items reflecting moral reasoning levels are in the upper .70s and low .80s. DIT scores are significantly related to cognitive capacity measures of moral comprehension (r= .60), to the recall and reconstruction of post-conventional moral arguments, to Kohlberg's measure of moral-development stages, and to other cognitive-developmental measures (Center for the Study of Ethical Development).

As described earlier, the pre-conventional stages of moral development involve egocentric attitudes and approaches to moral issues. They feature relatively immature moral motivations such as punishment avoidance, self-justification and gratification as an end in itself. Correspondingly, companion robots that encounter such expressions from humans they are charged to care for may either, depending on the construction of algorithms determining machine-learning processes, reinforce such lower-order moral claims or produce responses designed to encourage deeper reflection on them. Responses to expressions of despair or depressive states might suggest reflection on root causes, for example. Responses to expressions of self-harm or the potential of harming others might redirect focus from individual-level grievances to broader implications of well-bring and feelings that violence is a justifiable response to problems. Post-conventional, or higher-order moral reasoning, shifts the focus of moral deliberation from the self to social and community obligations. Whereas pre-conventional reasoning emphasizes utility ("How would this affect me?" "What might I get out of doing/not doing X?"), post-conventional deliberation is driven by broader priorities of Kantian duties that ensure respectful treatment of others, as well as Aristotelian questions about what it means to be socially responsible and how behaviors might promote or undermine flourishing for all. Items in the DIT reflecting higher stages of moral development, for example, include emphasize questions that transcend concerns of individual-level benefit involved in the scenarios, and shift focus to social implications and precedents of honoring obligations to the broader human community. For example, a DIT scenario involves a question of hiring an auto mechanic whose Chinese ethnicity might drive away prejudiced customers. Among the items reflecting higher-order moral reasoning include:

- "What individual differences ought to be relevant in deciding how society's rules are filled?"
- "Whether hiring capable men such as Mr. Lee would use talents that would otherwise be lost to society."

Another scenario features a terminal, suffering patient who begs her doctor to provide enough morphine to allow her to overdose. Among the items reflecting higher-order moral reasoning include:

- "Does the state have the right to force continued existence on those who don't want to live?"
- "Can society allow suicides or mercy killing and still protect the lives of individuals who want to live?"

Note that no items are intended to "guide" subjects to a particular conclusion, but reflect the essence of ethics: what matters is not the final decision made, but the level of informed reasoning and the prioritized values that define the deliberation. Similarly, rather than assigning "judgment" statements to ranges of subject expressions, machine-learning algorithms for companion units should weight expressions on an individualistic-moral/social obligation spectrum with appropriate responses that focus on qualities featured on one end of that spectrum and not the other. Algorithmic responses should emphasize human dignity and flourishing with statements that ask subjects to talk about perceived broader effects of contemplated actions – similar to the language of the DIT items above.

Machine learning and virtue theory

The project has aimed to introduce moral-development theory to the realm of machine-learning algorithm construction, and argue for its usefulness in considering the range of interaction goals for companion units. Language associated with higher-order moral reasoning emphasizes notions of universal respect and dignity and community belonging and obligation. It is premised on an expansive vision of justice. As such, it arguably draws less from a consequentialist, or utilitarian, ethical framework, and emphasizes principles from deontological and Aristotelian ethics. This emphasis will necessarily vary by machine-learning functionalities; in the realm of AI and self-driving vehicles, for example, it is difficult to envision what algorithms aimed at promoting a sense of Kantian duty or cultivating virtuous characters would look like. The aim, instead, is rightly the justifiable calculation of outcomes: how exactly AI-equipped vehicles would be programmed to prioritize drivers, passengers, or bystanders in collision scenarios, and execute emergency maneuvers accordingly, for example. But for machine-learning programs likely to confront less concrete situations, such as despairing, anxious or depressed patients, weighing various possible "outcomes" is arguably less central. The emphases on relations, connections, duties and virtues found in moral-development theory offer a more compelling foundation for robot companionship. The emphasis on virtue ethics frameworks, especially, comports with trends in moral psychology research (including studies on moral-reasoning) more broadly speaking: The study of the links among identity, moral motivation, theories of the self, value internalization, ethical ideology, and a host of other components of the moral self lends itself to a neo-Aristotelian framework emphasizing cultivation of character and embodiment of the virtues. Moral psychology scholar and bestselling author Jonathan Haidt and his colleagues have written eloquently on this point. While they challenge key assumptions of predominant methodologies (including use of the DIT) in moral-reasoning research, they identify the affinities between moral psychology research in general and virtue ethics frameworks:

> Part of the appeal of virtue theory has always been that it sees morality as embodied in the very structure of the self, not merely as one of the activities of the self....We believe that virtue theories are the most psychologically sound approach to morality. Such theories fit more neatly with what we know about moral development, judgment, and behavior (Haidt & Joseph, 2004, p. 61, 62).

As we enter an era in which machines are increasingly being assimilated into all areas of human life, more attention must be given to the way they will be "talking" to us. The development of algorithms for robot companions should be informed by narratives of morality – not necessarily to presume that robots have a moral life themselves, but to encourage moral thoughtfulness when providing conversational care to patients considering moral aspects of their own lives.

References

Bertolini, A., & Aiello, G. (2018). Robot companions: A legal and ethical analysis. *The Information Society, 34*(3), 130-140. doi:10.1080/01972243.2018.1444249

Boddington, P. (2019, March 21). Moral technology. *Aeon*. Available: https://aeon.co/essays/what-are-the-values-that-drive-decision-making-by-ai

Bogosian, K. (2017). Implementation of moral uncertainty in intelligent machines. *Minds and Machines, 27*(4), 591-608. doi:10.1007/s11023-017-9448-z

Brodwin, E. (2018, January 30). I tried Woebot, a therapy chatbot and app for depression. *Business Insider.* Retrieved from https://www.businessinsider.com/therapy-chatbot-depression-app-what-its-like-woebot-2018-1

Business Insider. (2015, May 13). Growth statistics for robots market 2015. *Business Insider*. Retrieved from https://www.businessinsider.com/growth-statistics-for-robots-market-2015-2?r=UK&IR=T

Centers for Disease Control and Prevention and National Association of Chronic Disease Directors. (2008). *The state of mental health and aging in America issue brief #1: What do the data tell us?.* National Association of Chronic Disease Directors. Retrieved from https://www.cdc.gov/aging/pdf/mental_health.pdf

Char, D.S., Shah, N.H., & Magnus, D. (2018). Implementing machine learning in health care: Addressing ethical challenges. *New England Journal of Medicine 378*(11), 981-983.

Coeckelbergh, M. (2017). *Using words and things: Language and philosophy of technology*. New York, NY: Routledge.

Coleman, R., & Wilkins, L. (2009). The moral development of public relations practitioners: A comparison with other professions and influences on higher quality ethical reasoning. *Journal of Public Relations Research, 21*(3), 318-340. doi:10.1080/10627260802520462

Fiske, S.T., & Taylor, S.E. (1984). *Social cognition*. Reading, MA: Addison-Wesley.

Guizzo, E. (2010, April 23). Hiroshi Ishiguro: The man who made a copy of himself. *Spectrum*. Retrieved from https://spectrum.ieee.org/robotics/humanoids/hiroshi-ishiguro-the-man-who-made-a-copy-of-himself

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Dædalus*, Fall 133(4), 55–65.

Home Health Aides and Personal Care Aides. (n.d.). *Bureau of Labor Statistics*. Retrieved from https://www.bls.gov/ooh/healthcare/home-health-aides-and-personal-care-aides.htm#tab-6

Isaac, A. M. C., & Bridewell, W. (2017a). White lies on silver tongues: Why robots need to deceive. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 157-172). New York, NY: Oxford University Press.

Ishiguro, H. (2006). Android science: Conscious and subconscious recognition. *Connection Science, 18*(4), 319-332. doi:10.1080/09540090600873953

Kohlberg, L. (1971). FROM IS TO OUGHT: How to commit the naturalistic fallacy and get away with it in the study of moral development. In T. Mischel, National Science Foundation (U.S.), & State University of New York at Binghamton (Eds.), *Cognitive development and epistemology* (pp. 151-235). New York: Academic Press.

La Bossiere, M. (2017b). Testing the moral status of artificial beings: or “I’m going to ask you some questions…”. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 293-306). New York, NY: Oxford University Press.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*(3), 297-337. doi:10.1075/is.7.3.03mac

Molteni, M. (2017, March 17). Artificial intelligence is learning to predict and prevent suicide. *Wired*. Retrieved from https://www.wired.com/2017/03/artificial-intelligence-learning-predict-prevent-suicide/

Mori, M. (2012, June 12). The uncanny valley: The original essay by Masahiro Mori. *Spectrum*. Retrieved from https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley

Parker, R.J. (1990). The relationship between dogmatism, orthodox Christian beliefs, and ethical judgment. *Counseling and Values, 34*(3), 213–216.

Piaget, J. (1932). *The moral judgement of the child*. London: K. Paul, Trench, Trubner.

Rest, J. (1979). *Development in judging moral issues*. University of Minnesota Press.

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (2000). A neo-Kohlbergian approach to morality research. *Journal of Moral Education, 29*(4), 381-395. doi:10.1080/713679390

Robinson, A. (2015, September 17). Meet Ellie, the machine that can detect depression. *The Guardian*. https://www.theguardian.com/sustainable-business/2015/sep/17/ellie-machine-that-can-detect-depression

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems, 86*, 152-161. doi:10.1016/j.robot.2016.08.018

Sabanovic, S., Bennett, C. C., Chang, W., & Huber, L. (2013). PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. *Paper presented at the 2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR), 2013,* 1-6. doi:10.1109/ICORR.2013.6650427

Simon, M. (2017, August 8). Companion robots are here. Just don’t fall in love with them. *Wired*. Retrieved from https://www.wired.com/story/companion-robots-are-here/

Smith, A., & Anderson, A. (2017). Americans’ attitudes toward robot caregivers. *Pew Research Center: Internet & Technology*. Retrieved from http://www.pewinternet.org/2017/10/04/americans-attitudes-toward-robot-caregivers/

Statista. (n.d.). Value of social and entertainment robot market worldwide from 2015 to 2025 (in billion U.S. dollars). *Statista*. Retrieved from https://www.statista.com/statistics/755684/social-and-entertainment-robot-market-value-worldwide/

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2017). *Report of COMEST on Robot Ethics*. Available: https://unesdoc.unesco.org/ark:/48223/pf0000253952

United States Census Bureau (2018). Older People Projected to Outnumber Children. *United States Census Bureau*. Retrieved from https://www.census.gov/newsroom/press-releases/2018/cb18-41-population-projections.html

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

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology, 12*(3), 243-250. doi:10.1007/s10676-010-9232-8

West, L.A., Cole, S., Goodkind, D., & He, W. (2014). *65+ in the United States 2010*. United States Census Bureau. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2014/demo/p23-212.pdf