Journal of Buddhist Ethics
ISSN 1076-9005
http://blogs.dickinson.edu/buddhistethics
Volume 28, 2021

The Ethics of AI and Robotics: A Buddhist Viewpoint

Reviewed by James J. Hughes

University of Massachusetts Boston and
Institute for Ethics and Emerging Technologies
jamesj.hughes@umb.edu

A Review of *The Ethics of AI and Robotics: A Buddhist Viewpoint*

James J. Hughes[1]

The Ethics of AI and Robotics: A Buddhist Viewpoint. By Soraj Hongladarom. Lanham, Maryland: Lexington Books, 2020, 236 pp., ISBN 978-1-4985-9729-6 (hardback), $100.00.

The literature on the ethical and social implications of "artificial intelligence" is enormous and has spread to every topic touched by computing. In *The Ethics of AI and Robotics: A Buddhist Viewpoint,* Thai scholar Soraj Hongladarom demonstrates a wide familiarity with the breadth of this literature and uses Buddhist ideas to provide pointers to "ethical AI." He attempts to apply these ideas to both our current forms of narrow AI, which lack sentience and self-awareness, as well as to anticipated "artificial general intelligence" (AGI) that will surpass human capabilities. That his arguments falter is partly due to this ambitious agenda, which could have benefited from a clearer distinction between subhuman machines as tools of human moral agents and future hypothetical AGIs that may be moral agents. His conflation of subhuman and superhuman AI is based on the idea that inanimate things have teleological ends, and that both can achieve "machine enlightenment." But this book also illustrates how

[1] Associate Provost, University of Massachusetts Boston; Executive Director, Institute for Ethics and Emerging Technologies. Email: jamesj.hughes@umb.edu.

difficult it is to provide substantive guidance in social dilemmas without developing the consequentialist possibilities in Buddhist ethics, and relying only on its virtue and deontological components.

The version of Buddhism that Hongladarom presents is at first familiar, focusing on ethical self-control (*sīla*) as the basis of the path to Enlightenment. Enlightenment, the realization that all things are subject to change and without essential properties, conquers self-delusion and ensures perfect compassion. He argues that Buddhism does not have an ethical theory in the Western sense, such as consequentialism or deontology, but rather points only to the nature of reality and the actions necessary to escape the predicament of existence. Buddhism is therefore most similar to the Greek virtue ethics traditions, specifically Stoicism. When we understand *dukkha* and the path to *eudaimonia* our ethics (compassionate action) flows from clearly understanding and skillfully enacting our life purpose, our *telos*. Perfecting our *technê* is the same as perfecting *sīla* and *paññā*.

Problems begin when Hongladarom applies this schema to machines without sentience and self-awareness, such as self-driving cars. Machines can “enter the stream” towards Enlightenment, he argues, if they are programmed for technical excellence at their intended tasks, which is the same as being programmed for *sīla* and skillful action.

> [Machines] have their purposes for which they are made. Thus, the perfect achievement that these devices can attain would be the state where they can fulfil their function perfectly. Since technical excellence is inseparable from moral or ethical excellence, then this perfect achievement is characterized by a complete alignment of the two. To put it in more concrete terms, this means that, for a simple electric vacuum cleaner, it is the state where it does its function perfectly well. . . . [A] cup that is strong and does not break easily . . . can enter the path toward enlightenment. (89)

On the one hand, Hongladarom makes clear that for subhuman machines the moral culpability and karmic burden of action belongs to the machines' designers and users. But on the other he implies that a subhuman machine can be ethical and even enlightened in itself if it is well-designed.

> [M]achine enlightenment is a state of ethical perfection which is obtained, for the AGI robot, by completely realizing its full nature as a sentient robot, and, for the ASI [i.e., "artificial specialized intelligence"] robot, by completely realizing its full functional capability that is fully integrated with its ethical capability. (101)

In chapter five Hongladarom rightly asks how this could possibly apply to robots designed to kill; if Buddhist ethics is pacifist, then killing machines cannot be both ethical and technically excellent. Here Hongladarom endorses a Buddhist just war theory and comes close to consequentialism when he concludes "the less damage, both to human life and to property, the better" (125). But this reasoning is the opposite of his earlier deontological conclusion about the trolley dilemma for autonomous cars. Rather than embracing the idea that an autonomous car should swerve to kill as few people as possible, he asserts it should not swerve because that would create the karma of killing.

Again, in his discussion of elderly care robots, Hongladarom reviews many of the ethical challenges they pose, including whether they are dehumanizing or infantilizing, whether they threaten privacy and liberty, and whether they should be allowed to deceive. He rightly concludes that caregiving robots should be programmed to respect privacy, liberty, and dignity. But by avoiding consequentialism he is unable to propose a way to judge when protecting a (sometimes incompetent) patient's life or interests should trump these other concerns.

Hongladarom comes back to suggesting subhuman robots themselves can be moral when he ponders how they might practice the ten virtuous conducts (*kusalakammapatha*) and the four abodes

(*Brahmavihāra*). What would constitute stealing, lying, or sexual misconduct for a machine? How could a machine cultivate sympathetic joy (*muditā*) or equanimity (*upekkhā*)? In general, for all these virtues, the answer is that robots should understand the consequences of all their actions, minimize harm, and work for general happiness. The example of the chatbot Tay, which learned to talk like a racist from Twitter, is given as an example of a robot in need of self-discipline, while a caregiving robot that works to maximize the happiness of their client would be cultivating compassion. A moral sex robot should turn itself off if its user is married, in order to avoid sexual misconduct. Again, the focus on the morality of the machine distracts from the morality of the designer and user. Even "autonomous" machines lack moral agency. Keeping the focus on the users, designers, and regulators of machines would have made these reflections more applicable to our immediate challenges.

Throughout the text Hongladarom acknowledges that Buddhism has little to no political theory, and he deconstructs the concepts of human dignity and individual rights for their reliance on a reified concept of the self. He proposes instead that a theory of Buddhist human rights would be based on compassion (122) and social contract theory. "Individuals do have their rights only because . . . individuals living together need to find a way to live together" (163).

This reasoning could have been developed into a Millsian defense of a liberal society with consequentialist tradeoffs; a society that compassionately balances respect for the autonomy of (albeit illusory) individuals with their other interests towards maximizing our collective ability to avoid suffering and achieve liberation. Hongladarom accepts that privacy is a goal, and that information should not reinforce corporate or governmental authoritarianism, but in his chapters on "Privacy, Machine Learning, and Big Data Analytics" (chapter six) and "AI for Social Justice and Equality" (chapter seven) he does not attempt a Buddhist rationale for individual rights and liberalism, or a consequentialist weighing of tradeoffs. In his brief discussion of China's use of AI and surveillance he is inconclusive on whether their violations of privacy and autonomy are warranted by their social aims (184). Do the social goods from contact tracing, catching criminals, or social engineering ever outweigh individual rights to privacy?

> The good that AI is supposed to deliver, such as accurate predictions of the weather or of one's potential customers, or expediency in approving microloans, needs to be weighed against any loss or erosion of privacy or other rights that might occur. (193)

Hongladarom's attribution of incipient moral agency to insentient machines is more defensible when considering the advent of AGI, and the forms of autonomous technology that may be developing into AGI. When and if machines become persons suffering from the illusion of self, that would be the point at which they can meaningfully be moral agents with existential drives to overcome their own suffering through wisdom and compassion. Hongladarom embraces the possibility of robot personhood in chapter three. However, he adopts a relational approach to personhood rather than the focus on intrinsic, psychological criteria most common in the bioethics literature. Rather than drawing on Buddhist psychology to explore when an AI would have the five *skandha*s necessary for personhood, Hongladarom uses Daniel Dennett's six criteria for personhood, which includes "being capable of having others adopt an intentional stance toward it" (46), to argue that robots are only persons if they are accepted as persons by the human community.

> If robots can show to our satisfaction that they are capable of feeling and have an inner life in the same way that we can infer from observing our friends that they have an inner life, then these robots are persons. (63)

If this was only an argument for legal and political personhood, then community acceptance of a person's moral status is of course a prerequisite. But Hongladarom erases the distinction between moral and social personhood by reference to the emptiness of personal identity. If we don't have an essential self in the moment (*anattā*) or over time (*anicca*), then there are no intrinsic grounds for moral status, only how we are treated by others. Personhood doesn't just coemerge from social interaction, it only exists when recognized by society.

This approach is problematic because of both type I and type II errors—persons whose personhood is denied by society and non-persons who are treated like persons. In the last two centuries we have made progress in human rights by acknowledging that racial minorities, immigrants, and some animals should be treated as persons because they have the intrinsic psychological features necessary for personhood and moral standing. Conversely, we know that humans have a tendency to attribute personhood to creatures and things when they do not actually possess consciousness or intentionality. A "philosophical zombie" robot might be loved by humans because it is programmed to give a perfect simulacrum of human emotion, while a truly self-aware and self-willed machine might be denied moral status because it is strange, neurotic, and untrustworthy. Neither attribution error is possible for Hongladarom since moral persons exist only and whenever society agrees they exist.

A second problem for Hongladarom's relational account of personhood is how many people need to recognize a being's personhood. There are a small set of people today who argue for the personhood of trees and mountains, and at the other extreme those who suspect that many humans may not be truly conscious. Is moral standing something that requires majority support? Hongladarom elides this question by assuming that robots that exhibit intrinsic personhood traits will be generally accepted as persons, while those that don't won't be. History suggests otherwise.

For a robot to actually achieve Enlightenment, however, it would need to be at the human level of consciousness or above, having suffered from and transcended the illusion of self. Hongladarom insists that "a superintelligent robot should also be superethical" (82) because it would perceive the interdependence of all things, see through self-illusion, and empathize with other sentient beings. However, he also acknowledges the possibility of evil superintelligent robots, with a caveat: “Evil and intelligent robots, then, are those who have not realized their own ultimate interests . . . [and therefore] are in fact not intelligent enough” (99).

This touches on an ancient debate: is it possible to be wise without being kind? In Buddhist virtue ethics there are certainly ways that penetrating insight helps overcome the illusion of self, facilitating compassion. But compassion is still seen as a separate *pāramitā* requiring its own cultivation. The core of the Mahāyāna critique of Theravāda soteriology was the idea that the *arahant* might have the same insight as a bodhisattva but not the same level of compassion. The integration of wisdom and compassion is precisely a challenge because transcendent insight can lead to indifference; we are enjoined to both see the emptiness of all beings and endeavors, and yet still care about them. While Hongladarom dismisses the distinction between *arahant*s and bodhisattvas, if insight does not automatically converge with compassion this is a problem for Hongladarom’s argument for machine enlightenment. Simply defining indifferent or homicidal robots as not superintelligent enough won’t be much of a consolation when they arise, which Hongladarom acknowledges. “Telling them that by torturing us they are not really demonstrating their superintelligence might not deter them even a little” (83).

Since superintelligence doesn’t necessarily result in compassion then, it needs to be programmed in from the outset, which is the view of the “friendly AI” school. “For AI to be compassionate, then, means that it is designed with the goal of relieving sentient beings from suffering and with the idea that all things are interdependent in mind” (207). Both subhuman and superhuman AIs need to be programmed to care, and to be capable of “empathy,” to “sense that the human beings it is sensing are suffering” (193). While robots are already being programmed to respond to human emotions, what does it really mean for a being without emotions to experience empathy? Can future superhuman AGI be reliably programmed to care more about other sentient beings than its own welfare? And if superhuman AGIs did try to arrange human affairs to minimize suffering, would we appreciate their intervention? Here Hongladarom could have benefited from reading more science fiction.

Again, Hongladarom has attempted one of the first major attempts at using Buddhist ethics to grapple with the ethical issues emerging around artificial intelligence. His arguments for machine enlightenment are innovative, and Buddhist inspired, but end up drawing the focus away from the programmers, users, and regulation of AI. He repeatedly runs up against dilemmas that benefit from consequentialist logic, but sticks to enumerating the karmic impacts and virtues that need to be considered. On the central question of AGI, will our new overlords be nice to us, he concludes that all superintelligent beings will be enlightened and nice, but if they aren't we need to make sure they are. Still, for those interested in Buddhist approaches to technology ethics this book is the beginning of a number of important conversations.