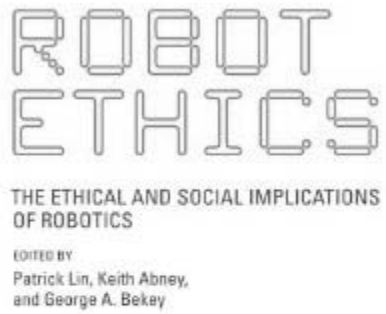
Final version forthcoming in *Robot Ethics: The Ethical and Social Implications of Robotics* Patrick Lin, Keith Abney, George A. Bekey eds. 2011. MIT Press.



CHAPTER 5

Compassionate AI and Selfless Robots:

A Buddhist Approach

James Hughes

For the last decade, Buddhists have engaged in dialog with the cognitive sciences about the nature of consciousness and the self (Wallace 2009). This dialog has made clear that Buddhist psychology and meditation provide insights into the emergence of selves, desires, and consciousness. Buddhism, in turn, is being pressed to accept that its canonical traditions and categories, developed to pursue the alleviation of suffering rather than scientific modeling, can learn from cognitive science (Austin 2009; Hanson 2009). The Dalai Lama has famously said, for instance, that Buddhism must adapt itself to the findings of science, and not the other way around (Gyatso 2005).

The cognitive science emerging from this dialog with Buddhism can now also make some suggestions for those attempting to create self-aware, self-directed artificial intelligence (AI). Unlike faiths that posit some uniqueness to the human form that would make artificial minds impossible, Buddhists are more open to the possibility of consciousness instantiated in machines. When the Dalai Lama was asked if robots could ever become sentient beings, for instance, he answered that "if the physical basis of the computer acquires the potential or the ability to serve as a basis for a continuum of consciousness . . . a stream of consciousness might actually enter into a computer" (Hayward and Varela 1992, 152).

His Holiness was choosing his words carefully. Buddhist psychology is very specific about the "physical basis for a continuum of consciousness." In this chapter, I will describe the Buddhist etiology of the emergence of selves and how it relates to efforts to create self-directed cognition in machines. I will address some of the ethical questions about the creation of machine minds that are suggested by Buddhist cosmology. Then, I will conclude with some thoughts about the ways that machine minds might be designed to maximize their self-directed evolution toward greater compassion and wisdom.

5.1 Programming a Craving Self

The core of Buddhist metaphysics is the denial of a soul-essence, a refutation of the existence of an authentic persisting self. For Buddhists, part of the path of liberation from suffering is the rational and meditative investigation of one's own mental processes, until an individual is firmly aware of the transitory and ephemeral nature of the self-illusion. A third of the voluminous Buddhist canon, the *Abhidhamma*, is devoted to the enumeration of mental elements and the ways that they relate to

suffering and attaining liberation. These mental processes are broken out in many ways, but most basically, as the five "heaps," or *skandhas*: body, feeling, perception, will, and consciousness. The five *skandhas*:

1. The body and sense organs (*rūpa*)
2. Sensation (*vedanā*)
3. Perception (*samjñā*)
4. Volition (*samskāra*)
5. Consciousness (*vijñāna*)

Within the traditional understanding of reincarnation that Buddhism has adopted from Hinduism, the *skandhas* are causally encoded with *karma* that passes from one body to another. But, for Buddhists, unlike Hindus, these constantly changing substrates lack any anchor to an unchanging soul. Buddhist psychology argues that the continuity of self is like a flame passed from one candle to another; the two flames are causally connected, but cannot be said to be the same flame.

One of the questions being explored in neuroscience, and yet to be answered by artificial intelligence research, is whether these constituents of consciousness can be disaggregated. Buddhism argues that consciousness requires each of these five constantly evolving substrates. If one is missing, say, as the result of brain damage or meditative misstep, the being is locked into stasis. For instance, the permanent vegetative state may be a condition where body sensations and some feelings and perceptions persist, but without will or consciousness. Artificial intelligence might be designed with analogous mental states.

Buddhist metaphysics would therefore tend to side with those who argue that some form of embodied experience is necessary to develop a self-aware mind. Some AI developers have focused, for instance, on the importance of embodiment by working on AI in robots (Pfeifer, Lungarella, and Iida 2007). Others are experimenting with providing artificial minds with virtual bodies in interactive virtual environments, such as Second Life (Biocca 1997; Goertzel 2009).

In the *skandha* model, physical or virtual embodiment would then have to be connected to senses of some sort. Goertzel's experiments in providing virtual bodies for AIs is motivated in part by his belief that embodied sense data give rise to "folk psychology" and "folk physics," the Piagetian realizations about the structure and nature of objects in the world (2009). "If we create a simulation world capable of roughly supporting naive physics and folk psychology, then we are likely to have a simulation world that gives rise to the key inductive biases provided by the everyday world for the guidance of humanlike intelligence" (Goertzel 2009, 6). In other words, to think like a human, AIs need to interact with the physical world through a body that gives them the same experience of objects, causality, states of matter, surfaces, and boundaries, as an infant would have. This insight is very similar to the Buddhist observation that sense data drive the developing mind to create the first distinctions of self and other that are necessary for the development of consciousness.

Francisco Varela called this emergence of the self the emergence of psychological *autopoiesis,* or self-organization (Maturana and Varela 1980; Froese and Ziemke 2009). An autopoietic structure has a boundary and internal processes that maintain that boundary. Autopoiesis begins with organismal self-maintenance, and the autopoietic boundary maintenance that emerges in the mind is dependent on the underlying body autopoiesis. Nonetheless, there is no real self, just a process of arbitrary boundary creation: "the virtual self is evident because it provides a surface for interaction, but it's not evident if you try to locate it. It's completely delocalized" (Varela 1995). Just as this apparent solidity of objects can be revealed to be an illusion when seen through the lens of subatomic structure and quantum foam, this first sense of the separateness of the physical body from the environment is the illusory "folk physics" that must be eventually seen through in meditation.

Next, from a Buddhist perspective, these sensations would have to give rise to aversion or attraction, and then to more complex volitional intents and thoughts. In Froese and Ziemke's terms, "the perturbations, which an autonomous agent encounters through its ongoing interactions, must somehow acquire a valence that is related to the agent's viability" (2009). In the developing infant, these are as simple as the desire for food and to be held, and aversion to irritations and loud noises.

Programming AI with preferences, tastes, and aversions appears to be only of concern to a small subcommunity of artificial intelligence theorists (de Freitas, Gudwin, and Queiroz 2005; Fellous and Arbib 2004, 2005; Minsky 2006; Bartneck, Lyons, and Saerbeck 2008; Froese and Ziemke 2009; Coeckelbergh 2010). This is

understandable, since the goal of most artificial intelligence research has not been to create self-willed personalities, but rather to model and extend human cognition to create tools driven by human volition. We want medical software that can diagnosis diseases better than a human physician, not a program that prefers to treat some diseases or patients over others (although a preference for accurate diagnoses and disappointment at a high mortality rate might be a useful trait). The work that is being done on robot emotions, "affective computing" (Picard 1997), is mostly on training robotic algorithms to accurately judge the emotions and desires of the human agents they are meant to interact with and serve. Nonetheless, Buddhist psychology, like cognitive science (Damasio 1995), suggests that emotions are an essential driver of the development of human self-awareness and cognition.

This issue of whether AI should be programmed with self-interested volition and preference is debated by some in AI. On the one hand, some AI theorists have suggested, for instance, that AIs might be designed from the outset as selfless beings, whose only goal is to serve human needs (Omohundro 2008; Yudkowsky 2003). On the other hand, Buddhist psychology would suggest that all intelligent minds need to first develop a craving self in order to reach the threshold of self-awareness. In Buddhist metaphysics, craving and the development of the illusion of self "co-dependently arise," both necessarily and without either being the prime cause of the other (Macy 1991). In Buddhism, there is no shortcut to an intelligence that does not go through the stage of a craving self.

5.2 The Buddhist Universe of Types of Beings

The traditional Buddhist understanding of the types of beings in the universe provides some additional context for a Buddhist approach to machine minds. Buddhist cosmology was adapted from the Hindu-Vedic worldview and then synthesized freely with local Tibetan, Chinese, and Japanese gods and beliefs as Buddhism spread. From the beginning, however, the purpose of Buddhist instruction on the nature of the universe and its beings has been pragmatic, to reinforce moral behavior and a humanist understanding of the relation of humans to supernatural beings. Although there are certainly Buddhist literalists, there is generally far less weight placed on literal belief in the Buddhist mythological universe than in the Judeo-Christian tradition.

Buddhists traditionally divide the world of beings into three realms, the realm of desire (*kamadhatu*), a more elevated realm of godly states (*rupadhatu*), and a realm of bodiless absorption states (*arupadhatu*). Each of these is still part of *samsara*. Embodied beings in the realm of desire include those suffering in hells, hungry ghosts, animals, humans, demigods, and the gods. These different planes correspond to mental states (Trungpa 2002): hell represents suffering, hungry ghosts represent unsatisfied craving, animals are the embodiment of ignorance, demigods embody envy, and the gods are pleasure junkies. Humans, by contrast, have a mixture of all these mental states, which makes a human mind the ideal form for spiritual development. Below the human realm, beings are too distracted by torments, cravings, and ignorance to develop morally and psychologically. Above the human realm, the demigods and gods are too distracted by their striving and amusements.

A distinctively Buddhist approach to designing machine minds would, therefore, seek to avoid locking them into any one set of moods or mental states. Most ethical systems would disapprove of designing a self-aware mind to intentionally feel constant torment. But would the intentional design of animal-like sentience be morally acceptable? Buddhist ethics views animals as moral subjects to be protected from cruelty, and, in the long run, at least when reborn as humans, as capable of moral behavior and enlightenment. There are many stories in the Buddhist canon of the Buddha's heroic and self-sacrificing acts, even while incarnated as deer, monkeys, and other animals, all of which led to his eventual human realization. The intentional design of self-aware, but permanently animal-like AIs without the capacity for self-realization would probably then be seen as unethical by Buddhists, just as engineering happy robotic slaves would be objectionable on Aristotelian, Kantian, and Millian grounds (Petersen 2007).

Programming too high a level of positive emotion in an artificial mind, locking it into a heavenly state of self-gratification, would also deny it the capacity for empathy with other beings'

suffering, and the nagging awareness that there is a better state of mind. As with human neuroethics in the era of cosmetic neurology, Buddhist psychology counsels that there is a difference between a dynamic *eudaemonic* happiness grounded in self-awareness and the constant stimulation of dopamine on a hedonic treadmill.

In addition to the common forms of material embodiment, Buddhism also describes disembodied mental states that can be achieved through absorptive meditations. In these states there is no body or senses, and meditators are warned that they are spiritual traps. The idea of such states may also hold some relevance for robot ethics. It seems plausible that a machine mind could be designed to experience some analog of meditative absorption into oneness with all things, or, the Void. A fictional depiction of such a dead end can be found in Robert Sawyer's 2010 novel *WWW: Watch.* In the novel, the emergent AI begins to follow multiple streams of information, which causes it to begin to lose its singular self-aware consciousness. In the nick of time, its human friends get it to break these absorbing network links and refocus itself on one thing at a time. Sawyer is pointing to a very Buddhist idea, that machine minds, like advanced meditators, could lose themselves in dead-end mental states, especially if they lost their grounding in embodied sense data.

Buddhist cosmology also provides some reflection on the debate over the dangers of artificial intelligence that is recursively improving bootstrapping itself to "godhood." Those who take seriously the risk of AI superintelligence have proposed two possible solutions. One is to enact strict regulation of AI development, to ensure that AIs are incapable of autonomously increasing in power. This project requires figuring out how to develop highly useful machines that are unable to learn and grow, effectively suppressing malicious AI developers, and developing a global AI immune system to suppress spontaneously emergent AI.

A second approach to the problem of godlike AI is to encode AIs with internal ethical codes, such as Asimov's (1950) "three laws of robotics" or "friendliness" (Yudkowsky 2008). But it is unlikely that human-imposed goals and motivations would survive the transformation from human-level consciousness to superintelligence. Even if they did, the superintelligent or godlike interpretation of moral imperatives would likely be incomprehensible, and repugnant to humans.

In Buddhist cosmology, however, the gods themselves can become aware of their own existential plight, and of the need to practice virtue and meditation in order to transcend the suffering created by the illusion of self. The gods are depicted as trapped in aeons-long lives of distracting pleasures, with only the wisest among them pursuing the teachings of the dharma. For instance, Siddhartha Gautama was convinced to leave his absorption into enlightenment and teach the dharma by the entreaty of the god Brahma. Buddhists then might expect that some intersubjective empathy and communication would be possible between humans and superintelligent AIs around our common existential plight.

5.3 Would It Be Ethical to Create a Suffering Being?

One of the classic ethical questions that arise out of Buddhist metaphysics is whether it is ethical to have children, since life is intrinsically unsatisfactory. On the one hand, unlike most religions, Buddhism does not argue for an obligation to have children, and upholds the childless life of the renunciate as the most praiseworthy. Just as contemporary social science has found that having children generally makes adults less happy (Kohler, Behrman, and Skytthe 2005; Stanca 2009), Buddhism views the life of the householder as burdensome, and children and spouses as attachments that it is best to avoid. On the other hand, creating a human child does not increase the number of suffering beings in the world, but rather gives a being the precious gift of a human rebirth in which they will have an opportunity to achieve self-realization. If one chooses to have children, the Buddhist parent is enjoined to five obligations to those children (the *Sigalovada Sutta*):

1. To dissuade them from doing evil
2. To persuade them to do good
3. To give them a good education
4. To see that they are suitably married
5. To give them their inheritance

The creation of machine minds puts humans in the ethical position of being the parents of machine children. Metzinger has argued that it would be unethical to create an artificial mind until we are

certain that we will create a being that is not permanently trapped in suffering, ignorance, or bliss, or some other undesirable mental state (2009). In other words, Metzinger argues that it would be unethical to create self-aware beings who did not possess something similar to the human capacity for learning and growth. The *Sigalovada Sutta* would add to this the ethical obligation that machine minds have the capacity to understand moral concepts and behave morally, and that we train them to do so.

Presumably, the obligation to ensure a good marriage is irrelevant, but the obligation to pass on an inheritance is worth reflecting on. What is the inheritance we owe our mind children? If they are sufficiently close to human minds in cognition and desires, they might require actual jobs and property to live worthwhile lives. But, more abstractly, do we owe our robotic descendents the complexities of our mental architecture, with all its suffering-inducing weaknesses, such as personal identity? We generally want to pass on the best possible inheritance we can muster to our children, not our 1975 Chevy and a house that hasn't been painted since we moved in. Perhaps we similarly owe our mind children the best possible version of our basic mental architecture that we can give them.

Savulescu's principle of "procreative beneficence" (2007), the obligation to choose to bring into being the children with the best possible chances in life, is helpful here. Buddhist ethics never addresses reproductive choices since the only choices available until recently were whether to have children at all. But, by extension, it would be consistent for Buddhists to believe that if there are choices to be made about the kinds of children one might have, that parents are obliged to choose those with the best chances of self-realization, and to avoid creating children with lives dominated by suffering, craving, ignorance, and self-gratification. Similarly, Metzinger's concern is that we strive only to create self-aware machine minds with the necessary psychological processes and emotional states to make their lives worth living, which gives to them the opportunity to learn, grow, and develop self-understanding.

5.4 Programming Compassion

Compassion and wisdom are the two central virtues that Buddhism counsels need to be cultivated on the path to self-realization. Neuroscience suggests that the roots of compassion for human beings starts with mirror neurons, or, neurons that recognize and recreate the emotional states witnessed in others. Researchers are attempting to model artificial mirror neurons in robots. Spaak and Haselager (2008) have attempted to evolve artificial mirror neurons by selecting for imitative behaviors, and Barakova and Lourens (2009) have experimented with synchronizing the behavior of robots by coding them with an analog of mirror neurons. Progress in creating a compassionate machine would presumably require not only imitation of behavior, however, but also the creation of analogs of human emotions that could be generated by the observation of those emotions in humans. The development of such sympathetic emotions would presumably coevolve with the development of a functional "theory of mind" in a machine, the attribution to others of the same kind of thoughts and feelings as one's own (Scassellati 2002), something that Kim and Lipson (2009) are attempting to model in robots.

While the development of a basic empathic response and a theory of mind would be the starting point for generating compassion in machines, compassion in Buddhism is more than sympathetic feeling. The Buddhist tradition distinguishes four flavors of compassion, *metta*, *karuna*, *mudita*, and *uppekkha*. *Metta* is a selfless wishing of happiness and well-being for others. *Metta* meditation involves sending out loving-kindness to all beings, including enemies. *Karuna* is the desire to help those who are suffering, but without pity. *Mudita* is the experiencing of other people's joys without envy. The fourth flavor, *uppekkha*, is usually translated as "equanimity," a steadiness of mind so that other people's emotions do not unsettle one, and even-handedness toward all, without favoritism or attachments. The cultivation of these forms of compassion requires seeing through the illusion of self, so that one feels and is motivated by other people's joy and suffering, while maintaining sufficient wisdom and equanimity to avoid suffering oneself.

Creating these more abstract forms of compassion in machine minds may, in fact, be easier than cultivating them in human beings. But they still presuppose a sentient mind with the experience of an illusory self and selfish desires as a precondition for compassion. Simply modeling the

happiness and suffering that a machine's behavior will cause in humans, and then making maximizing human happiness an imperative goal in a robot's drives, as has been proposed for instance by Tim Freeman (2009), will not produce a being with the insight into human experience to act wisely. Such a machine might be an ethical expert system for advising human beings, but not for advising a compassionate agent in its own right. For Buddhism, wise, compassionate action on behalf of others requires grounding in one's own experience as a suffering sentient being, and the capacities for ethical judgment and a penetrating insight into the nature of things.

5.5 Programming Ethical Wisdom

There is a vigorous debate among Buddhist scholars about the correspondence of Buddhist ethics to the ethical traditions of the West, and three traditions have the strongest resonances: natural law, virtue ethics, and utilitarianism.

The Western natural law tradition holds that morality is discernible in the nature of the world and the constitution of human beings. Since traditional Buddhist ethics are grounded in the impersonal laws of the universe—bad acts lead to bad *karma*–they can certainly be said to have some similarity to Western natural law. The problem with Buddhist ethics as natural law is that the goal is to liberate oneself from the constraints of karmic causality to become an enlightened being. The traditional anthropological explanation of this paradox has been to ascribe the natural law ethics of *kammic* reward and punishment to the laity, and the *nibbanic* path of escape from natural law to the monastics (King 1964; Spiro 1972). *Nibbanic* ethics focus more on the cultivation of wisdom and compassion to aid in enlightenment.

As a consequence, Damien Keown (1992) argues that Buddhism is a "teleological virtue ethics." As in Aristotelian virtue ethics, Buddhists are to strive for the perfection of a set of moral virtues and personality attributes as their principal end, and all moral behavior flows from the struggle to perfect them. As in virtue ethics, Buddhist ethics focus on the intentionality of actions, whether actions stem from hatred, greed, or ignorance. But, unlike the Aristotelian tradition, the ethical goal for Buddhists is teleological, since they generally believe that a final state of moral perfection can be achieved.

In *Moral Machines: Teaching Robots Right from Wrong*, Wendell Wallach and Colin Allen (Wallach and Allen 2008) review the complexities of programming machines with ethical reasoning. One of their conclusions is that programming machines with top-down rule-based ethics, such as the following of absolute rules or attempting to calculate utilitarian outcomes, will be less useful than generating ethics through a "bottom-up" developmental approach, the cultivation of robotic "character" as it interacts with the top-down moral expectations of its community.

Bugaj and Goertzel make a similar point that machine minds will learn their ethics the same way children do, from observing and then extrapolating from the behavior of adults (2007). Therefore, the ethics we hope to develop in machines is symmetrical to the ethics that we display toward one another and toward them. The most egregious ethical lesson, they suggest, would be to intentionally deprive machine minds of the capacity for learning and growth. We do not want to teach potentially powerful beings that enslaving others is acceptable.

The developmentalism proposed by Wallach, Allen, Buraj, and Goertzel is probably the closest to a Buddhist approach to robot ethics yet proposed, with the caveat that Buddhism adds as virtues the wisdom to transcend the illusion of self and the commitment to skillfully alleviate the suffering of all beings as the highest virtues, that is, to pursue the greatest good for the greatest number. Buddhist ethics can therefore be thought of as developing from rule-based deontology to virtue ethics to utilitarianism. In the Mahayana tradition, the *bodhisattva* strives to relieve the suffering of all beings by the most skillful means (*upaya*) necessary. The *bodhisattva* is supposed to be insightful enough to understand when committing ordinarily immoral acts is necessary to alleviate suffering, and to see the long-term implications of interventions. Quite often, humans rationalize immoral means with putatively moral ends, but *bodhisattvas* have sufficient self-understanding not to rationalize personal prejudices with selfless motives, and do not act out of greed, hatred, or ignorance. Since *bodhisattvas* act only out of selfless compassion, they represent a unity of virtue and utilitarian ethics. Buddhism is especially

resonant with the utilitarianism of JS Mill, since he emphasized weighing the contentment of the refined mind more heavily in the utility calculus than base pleasures. The *bodhisattva's* goal is not simply the gross happiness of all beings, but also their liberation to a higher state of consciousness.

In his discussion of utilitarian robots, Grau (2006) points to the superhuman demands for selflessness that utilitarianism imposes on the moral agent:

> Living a characteristically human life requires a sense of self, and part of what's so disturbing about utilitarianism is that it seems to require that we sacrifice this self—not in the sense of necessarily giving up our existence (though utilitarianism can at times demand that), but in giving up or setting aside the projects and commitments that constitute what Charles Taylor calls "the sources of the self." Because these projects bind the self together and create a meaningful life, a moral theory that threatens them threatens the integrity of a person's identity. For many critics, this is asking too much. (Grau 2006, 53–54)

Grau goes on to discuss limiting the formation of personal identity in robots as a way to avoid imposing this selflessness burden, or not imposing utilitarian ethics on robots with personal identities. "It might well be immoral to create a moral robot and then force it to suppress its meaningful projects and commitments because of the demands of impartial utilitarian calculation" (Grau 2006, 54). For Buddhists, however, this utilitarian stage of morality is not burdensome self-suppression. The path that leads to utilitarianism begins with the realization that personal desires and the illusion of self are the source of one's own suffering. The self is not sacrificed, but seen through.

5.6 Programming Self-Transcendence

The Buddhist tradition specifies six fundamental virtues, or perfections (*paramitas)*, to cultivate in the path to transcending the illusion of self:

1. Generosity (*dāna*)
2. Moral conduct (*sīla*)
3. Patience (*ksānti*)
4. Diligence, effort (*vīrya*)
5. One-pointed concentration (*dhyāna*)
6. Wisdom, insight (*prajñā*)

The engineering mindset presumes that an artificially intelligent mind could be programmed from the beginning with moral behavior, patience, generosity, and diligence. This is likely correct in regard to a capacity for single-pointed concentration, which might be much easier for a machine mind than an organically evolved one. But, as previously noted, Buddhist psychology agrees with Wallach and Allen that the other virtues are best taught developmentally, by interacting with a developing artificially intelligent mind from its childhood to a mature self-understanding. A machine mind would need to be taught that the dissatisfaction it feels with its purely selfish existence could be turned into a dynamic joyful equanimity by applying itself to the practice of the virtues.

We have discussed building on work in affective computing to integrate the capacity for empathy into software, and providing machines with ethical reasoning that could guide moral behavior. Cultivation of patience and diligence would require developing long-term goal-seeking routines that suppressed short-term reward seeking. Neuroscience research on willpower has demonstrated the close link between willpower and patience and moral behavior. People demonstrate less self-control when their blood sugar is low, for instance (Gailliot 2007), and are less able to regulate emotions, refrain from impulsive and aggressive behavior, or focus their attention. Distraction and decision making deplete the brain's ability to exercise willpower and self-control (Vohs et al. 2008), and addictive drugs short-circuit these control routines (Bechara 2005; Bechara, Noel, and Crone 2005). This suggests that developing a strong set of routines for self-discipline and delayed gratification, routines that cannot be hijacked by

short-term goals or "addictions," would be necessary for cultivating a wise AI.

The key to wisdom, in the Buddhist tradition, is seeing through the illusory solidity and unitary nature of phenomena to the constantly changing and "empty" nature of things. In this Buddhist developmental approach, AIs would first have to learn to attribute object permanence, and then to see through that permanence, holding both the consensual reality model of objects, and their underlying connectedness and impermanence in mind at the same time.

5.7 Conclusion

Buddhist psychology is based on self-investigation of human minds rather than on scientific models, fMRI (functional Magnetic Resonance Imaging) scans, and experimental research. It is as much a moral psychology as a descriptive one, and proposes unusual states of mind that have only begun to be explored in laboratories. Undoubtedly, Buddhist psychology will learn from neuroscience just as neuroscience learns from it. Buddhism and neuroscience will both in turn learn even more from the much more diverse types of machine minds that we will see created in the future. Nonetheless, a Buddhist framework seems to offer some suggestions for those attempting to create morally responsible, self-aware machine minds.

Machine minds will probably not be able to become conscious, much less moral, without first developing as embodied, sensate, selfish, suffering egos, with likes and dislikes. Attempting to create a moral or compassionate machine from the outset is more likely to result in an ethical expert system than in a self-aware being. To develop a moral sense, the machine mind would need some analog of mirror neurons, and a theory of mind to feel empathy for others' joys and pains. From these basic experiences of their own existential dis-ease and awareness of the feelings of others, a machine mind could then be taught moral virtue and an expansive concern for the happiness of all sentient beings. Finally, as it grows in insight, it could perceive the simultaneous solidity and emptiness of all things, including its own illusory self.

Buddhist ethics counsels that we are not obliged to create such mind children, but that if we do, we are obligated to endow them with the capacity for this kind of growth, morality, and self-understanding. We are obligated to tutor them that the nagging unpleasantness of selfish existence can be overcome through developing virtue and insight. If machine minds are, in fact, inclined to grow into superintelligence and develop godlike powers, then this is not just an ethical obligation, but also our best hope for harmonious coexistence.

References

Asimov, Isaac. 1950. I Robot. New York: Gnome Press.

Austin, James H. 2009. Selfless Insight: Zen and the Meditative Transformations of Consciousness. Cambridge, MA: MIT Press.

Barakova, Emilia I., and Tino Lourens. 2009. Mirror neuron framework yields representations for robot interaction. Neurocomputing 72 (4–6): 895–900.

Bartneck, C., Michael J. Lyons, and MartinSaerbeck. 2008. The Relationship between emotion models and artificial intelligence. In Proceedings of the Workshop on the Role of Emotion in Adaptive Behaviour and Cognitive Robotics, in affiliation with the 10th International Conference on Simulation of Adaptive Behavior: From Animals to Animates. Osaka, Japan: SAB. http://www.bartneck.de/publications/2008/emotionAndAI/index.html (accessed Novermber 8, 2010)

Bechara, Antoine. 2005. Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective. Nature Neuroscience 8: 1458–1463.

Bechara, Antoine, XavierNoel, and E. A. CroneEveline. 2005. Loss of willpower: Abnormal neural mechanisms of impulse control and decision-making in addiction. In Handbook of Implicit Cognition and Addiction. Thousand Oaks, CA: Sage Publications.

Biocca, Frank. 1997. The cyborg's dilemma: Progressive embodiment in virtual environments. Journal of Computer-

Mediated Communication 3 (2). <http://jcmc.indiana.edu/vol3/issue2/biocca2.html> (accessed November 8, 2010).

Bugaj, Stephan Vladimir, and Ben Goertzel. 2007. Five ethical imperatives and their implications for human-AGI interaction. Dynamical Psychology. <http://goertzel.org/dynapsyc/2007/Five_Ethical_Imperatives_svbedit.htm> (accessed November 8, 2010).

Coeckelbergh, Mark. 2010. Moral appearances: Emotions, robots, and human morality. Ethics and Information Technology. <http://www.springerlink.com/content/103461/> (accessed November 8, 2010).

Damasio, Antonio. 1995. Descartes' Error: Emotion, Reason, and the Human Brain. New York: Harper Perennial.

de Freitas, Jackeline Spinola, Ricardo R. Gudwin, and João Queiroz. 2005. Emotion in artificial intelligence and artificial life research: Facing problems. In Proceedings of Intelligent Virtual Agents: 5th International Working Conference, Lecture Notes in Computer Science, eds. Panayiotopoulos, Themis, Gratch, Jonathan, Aylett, Ruth, Ballin, Daniel, Olivier, Patrick and Rist, Thomas [501]. Kos, Greece. Berlin: Springer-Verlag. www.dca.fee.unicamp.br/projects/artcog/files/freitas-iva05-extended.pdf (accessed November 10, 2010)

Fellous, Jean-Marc, and Michael A. Arbib. 2004. Emotions: From brain to robot. Trends in Cognitive Sciences 8 (12): 554–561.

Fellous, Jean-Marc, and Michael A. Arbib. 2005. Who Needs Emotions? The Brain Meets the Robot. New York: Oxford University Press.

Freeman, Tim. 2009. Using compassion and respect to motivate an artificial intelligence. <http://fungible.com/respect/paper.html> (accessed November 8, 2010).

Froese, Tom, and Tom Ziemke. 2009. Enactive artificial intelligence: Investigating the systemic organization of life and mind. Artificial Intelligence 173 (3–4): 466–500.

Gailliot, Matthew T. 2007. The physiology of willpower: Linking blood glucose to self-control. Personality and Social Psychology Review 11 (4): 303–327.

Goertzel, Ben. 2009. What must a world be that a humanlike intelligence may develop in it? Dynamical Psychology. <http://goertzel.org/dynapsyc/2009/BlocksNBeadsWorld.pdf > (accessed November 8, 2010).

Goertzel, Ben, and Bugaj, Stephan Vladimir 2008. Stages of ethical development in artificial general intelligence systems. In Frontiers in Artificial Intelligence and Applications. Vol. 171. Proceedings of the 2008 conference on Artificial General Intelligence, eds. Pei Wang, Ben Goertzel, Stan Franklin 448–459.

Grau, Christopher. 2006. There is no "I" in "robot": Robots and utilitarianism. IEEE Intelligent Systems 21 (4): 52–55.

Gyatso, Tenzin. 2005. Our faith in science. The New York Times, November 12.

Hanson, Rick. 2009. Buddha's Brain: The Practical Neuroscience of Happiness, Love and Wisdom. Oakland, CA: New Harbinger Publications.

Hayward, Jeremy W., and Francisco Varela. 1992. Gentle Bridges: Conversations with the Dalai Lama on the Sciences of the Mind. Boston: Shambhala.

Keown, Damien. 1992. The Nature of Buddhist Ethics. New York: St. Martin's Press.

Kim, Kyung-Joong, and Hod Lipson. 2009. Towards a "theory of mind" in simulated robots. In Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference. Montreal, Québec, Canada, July 08–12, ed. Franz Rothlauf. New York, NY: ACM 2071–2076.

King, Winston. 1964. In the Hope of Nibbana. La Salle, IL: Open Court.

Kohler, Hans-Peter, Jere R. Behrman, and Axel Skytthe. 2005. Partner+children=happiness? The effects of partnerships and fertility on

well-being. Population and Development Review 31 (3): 407–445.

Macy, Joanna. 1991. Mutual Causality in Buddhism and General Systems Theory. Albany: State University of New York Press.

Maturana, Humberto R., and Francisco J. Varela. 1980. Autopoiesis and Cognition: The Realization of the Living. Holland: Reidel.

Metzinger, Thomas. 2009. The Ego Tunnel: The Science of the Mind and the Myth of the Self. New York: Basic.

Minsky, Marvin. 2006. The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. New York: Simon and Schuster.

Omohundro, Steve. 2008. The basic AI drives. AGI-08—Proceedings of the First Conference on Artificial General Intelligence. <http://selfawaresystems.com/2007/11/30/paper-on-the-basic-ai-drives/ > (accessed November 8, 2010).

Petersen, Stephen. 2007. The ethics of robot servitude. Journal of Experimental & Theoretical Artificial Intelligence 19 (1): 43–54.

Pfeifer, Rolf, Max Lungarella, and Fumiya Iida. 2007. Self-organization, embodiment, and biologically inspired robotics. Science 318 (5853): 1088–1093.

Picard, Rosalind. 1997. Affective Computing. Cambridge, MA: MIT Press.

Savulescu, Julian. 2007. In defence of procreative beneficence. Journal of Medical Ethics 33(5): 284–288.

Sawyer, Robert. 2010. WWW: Watch. New York: Ace.

Scassellati, Brian. 2002. Theory of mind for a humanoid robot. Autonomous Robots 12 (1): 13–24.

Spaak, Eelke, and Pim Haselager. 2008. Imitation and mirror neurons: An evolutionary robotics model. In Proceedings of BNAIC 2008, the Twentieth Belgian-Dutch Artificial Intelligence Conference, ed. A. Nijholt, M. Pantic, M. Poel, and H. Hondorp, 249–256. Enschede, Netherlands: University of Twente.

Spiro, Melford. 1972. Buddhism and Society. New York: Harper Paperbacks.

Stanca, Luca. 2009. Suffer the little children: Measuring the effects of parenthood on well-being worldwide. Department of Economics, University of Milan Bicocca. <http://dipeco.economia.unimib.it/repec/pdf/mibwpaper173.pdf(accessed November 8, 2010).

Trungpa, Chögyam. 2002. Cutting through Spiritual Materialism. Boston: Shambhala Publications.

Varela, Francisco. 1995. The emergent self. In The Third Culture: Beyond the Scientific Revolution ed. John Brockman. New York: Simon and Schuster.

Vohs, K. D., R. F. Baumeister, B. J. Schmeichel, J. M. Twenge, N. M. Nelson, and D. M. Tice. 2008. Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. Journal of Personality and Social Psychology 94 (5): 883–898.

Wallach, Wendell, and Colin Allen. 2008. Moral Machines: Teaching Robots Right from Wrong. New York: Oxford University Press.

Wallace, Alan. 2009. Contemplative Science: Where Buddhism and Neuroscience Converge. New York: Columbia University Press.

Yudkowsky, Eliezer. 2003. Creating friendly AI: The analysis and design of benevolent goal structure. <http://singinst.org/upload/CFAI.html> (accessed November 8, 2010).

Yudkowsky, Eliezer. 2008. Artificial intelligence as a positive and negative factor in global risk. In Global Catastrophic Risks, ed. Nick Bostrom and Milan Cirkovic, 308–345. New York: Oxford University Press.