Machine Learning, Plant Learning, and the Destabilization of Buddhist Psychology

CHARLES GOODMAN
Binghamton University
cgoodman@binghamton.edu

Abstract: Recent developments in artificial intelligence and the nascent scientific literature on 'plant learning' pose serious challenges to Buddhist philosophy of mind and to Buddhist practical ethics. These challenges are of two general types. First, the empirical results threaten to extend the reach of mind more broadly than premodern South Asian and Tibetan Buddhists were willing to allow, calling into question the rational defensibility of a range of Buddhist moral commitments.

But the discovery of learning in non-animals also threatens to destabilize the crucial Buddhist distinction between 'sentient beings' and the 'receptacle world', and raises the possibility of a separation between intelligence and consciousness. The emergence of such a separation could require a basic rethinking of the traditional framework of the five aggregates. These developments should also sharpen our attention to AI safety by making the prospect of existential AI risk even more threatening than it would otherwise have been.

Keywords: AI safety, existential risk, Buddhist psychology, five aggregates, vegetarianism, plant minds

DOI: : https://dx.doi.org/10.15239/hijbs.03.02.03

Introduction

At the heart of Buddhist ethics as I understand it, is the aspiration to relieve the suffering and promote the welfare of all sentient beings. But to make any progress on such an intention, we will often need to know the answer to a demarcation question: of the various types of observable objects and systems in the world around us, which are the sentient beings? The Buddhist tradition, in its South Asian and Tibetan articulations at least, has usually believed itself to possess a correct and relatively simple answer to this practically and normatively important question. The sentient beings are the inhabitants of the six realms: humans, nonhuman animals, and various hard-to-observe beings such as hungry ghosts, titans and gods. Plants are definitively excluded from the cycle of rebirth and from the status of sentience; along with nonliving physical objects, they form the 'receptacle world', the stage on which the activities of sentient beings can unfold.

Recent developments in biology, robotics and computer science have been putting more and more pressure on this simple answer. Later in this paper, I will consider the profoundly disturbing implications of the developing technology of machine learning, both for the viability of Buddhist psychological theories and, in a very different way, for the future of the human race. First, though, I will discuss the possible implications of a small but growing literature claiming to find empirical evidence of learning in plants. I am neither a botanist nor an experimental psychologist, and it is far outside my expertise to evaluate this evidence. In view of the ongoing replication crisis in the behavioral sciences, some caution about it is certainly in order. But taken at face value, the experimental results can be read as providing significant new support to the Jaina position that plants are sentient beings. Along with interesting theoretical quandaries, these findings raise a very practical question: Now what do we eat?

I. Plant Learning and Its Implications

For years it has seemed to me that the most crucial question for assessing the moral standing of plants concerns whether they are capable of learning from experience. The argument for resting the weight on this point is quite straightforward and simple. In animals, we have abundant evidence that the central psychological role of pleasure and pain has to do with learning. In typical cases, an animal that experiences pain learns to avoid the situation or the entity that produced the pain; an animal that experiences pleasure learns to repeat the conditions that produced that sensation. Learning, as we understand it, requires such a reinforcement signal; so, if a living system learns, we have reason to believe that it experiences pleasure and pain, or at the very least, has mental states that play a similar functional role to pleasure and pain.

Now, there is an ongoing and, indeed, interminable debate about what features of human lives are intrinsically morally important. But most authors who have advocated the moral claims of nonhuman animals hold that what we have the most moral reason to attend to as regards them is that we not cause them pain. Some authors, notably Peter Singer, argue that the pain of nonhuman animals must be seen as just as important as human pain; others, such as Bonnie Steinbock, argue that it is less important;[1] but both sides would rest an objection against factory farming or inhumane laboratory practices precisely on this issue of pain.

Scientists have made truly remarkable discoveries about the complexities of plant behavior. But if the above argument is correct, most of these do not provide much support for thinking that plants have moral standing. So, for example, it has long been known that certain fungi supply plants with materials such as nitrates and phosphates from the soil, and that, in return, the plants send the fungi nutrients such as sugars and fats that are the products of photosynthesis. In a remarkable recent finding, biologists discovered that the rate at which the fungi exchange these materials varies with their scarcity in the

---

[1] See Steinbock, 'Speciesism and the Idea of Equality'.

environment.[2] This is an astonishing demonstration of the power of evolution, but on reflection, it should not move the needle much on the issue of moral standing. The behavior could easily have been implemented solely by a genetically programmed and wholly unconscious algorithm.

Learning may well be another matter. For decades, studies have suggested, in tantalizing but inconclusive ways, that plants might be capable of learning.[3] Recently, though, experimental results opened up the possibility of a genuinely convincing case for plant learning.[4] There are several of these studies, but the most prominent and perhaps the most convincing example is a series of experiments on garden peas performed by Monica Gagliano and collaborators and published in *Nature Scientific Reports*.[5] In one of these tests, Gagliano and her team exposed pea seedlings to a simple maze with a Y-bifurcation. Pea seedlings, like many other seedling plants, naturally grow towards the light, which they need for photosynthesis. The seedlings in the control group were presented with light coming from one of the two branches of the 'Y'; as expected, all of them grew towards that side of the maze. In the test group, on the other hand, the seedlings were 'trained' by presenting them with a light source and a fan. This group in turn was divided into an [F + L] group, for which the fan and the light source always came from the same branch of the maze, and an [F vs. L] group, for which the fan was placed in the opposite branch from the light. During the experiment, the test group seedlings were then exposed to the fan for three ninety-minute

---

[2] Whiteside, et al., 'Mycorrhizal Fungi Respond to Resource'. For a brief popular presentation see 'An underground marketplace', *Economist*, June 8, 2019, https://www.economist.com/science-and-technology/2019/06/06/fungi-it-turns-out-are-canny-traders-of-nutrients-to-plants.

[3] For a review of some of this older evidence, along with interesting proposals about how to theorize about plant minds, see Trewavas, 'Aspects of Plant Intelligence'.

[4] Thanks to Prof. Kristin Andrews of York University for calling this literature to my attention.

[5] Gagliano, et al., 'Learning by Association in Plants'.

periods, and then were checked the next morning to see where they had grown. As the scientists reported:

> in the test group, the majority of seedlings exhibited a conditioned response to the fan. In the [F + L] group, 62% of the seedlings grew towards the fan, whereas in the [F vs L] group, 69% of the seedlings grew in the direction opposite to the fan.[6]

Thus, the plants were able to learn by association. In fact, this learning had a more powerful effect on their growth behavior than did their natural tendency, dominant in the control group, to grow towards wherever the light previously appeared. Gagliano herself, in a short piece reflecting on the implications of her results, makes it clear that she thinks they are evidence for the existence of plant subjectivity, and even that they place the question of moral standing very much on the table for discussion.[7]

The idea that plants might have some form of sentience is not altogether alien even to South Asian articulations of Buddhism. Harvey assembled some interesting evidence for the proposition that early Buddhists may have thought of plants as 'one-facultied life' (Pāli *ekīndriya jīva*).[8] Some of this evidence is less impressive than it first appears. Harvey writes that 'after a reference to people's concern over "one-facultied life", the Buddha criticizes a monk who has cut down a large tree used as a shrine, saying, "For, foolish man, people are percipient of a life-principle in a tree"'.[9] The quote from the Buddha, in the original Pāli, is *jīva-saññino hi moghapurisa manussā*

---

[6] Gagliano, et al., 'Learning by Association in Plants', 2. References to graphical figures presented in the original are omitted in this quotation.

[7] Gagliano, 'The mind of plants', 3. Gagliano has also performed experiments that seem to provide evidence of learning in *Mimosa pudica*, the so-called 'sensitive plant'; but her interpretation of these results has led to controversy, and it is not yet clear exactly what they mean and whether they can be replicated. See Biegler, 'Insufficient evidence for habituation'.

[8] Harvey, *An Introduction to Buddhist Ethics*, 174–76.

[9] Harvey, *An Introduction to Buddhist Ethics*, 175, citing *Vin*. III.156.

rukkhasmim. But the Pāli term *saññā*, equivalent to Sanskrit *saṃjñā*, is better translated as 'conceive' than 'perceive'—as we can see from numerous passages from various stages of Buddhist scriptural literature in which the term is used in the context of telling practitioners to conceive of some X as being a Y. In English, 'perceive' is normally a success verb; at least if we are speaking at the conventional level of truth, if I perceive a table in this room, then there is a table. However, people conceive of things around them in all kinds of distorted ways. If, as I would suggest, the passage would be better translated as, 'For, foolish man, people conceive of a tree as containing a life-principle', then the wording no longer suggests in any way that the people are correct to conceive of it in that way. Such a translation makes more plausible an interpretation on which the motivation for the rule is simply to spare the feelings of those who do not share a Buddhist outlook.

Despite this issue, Harvey offered enough other textual evidence to motivate the conclusion that there was indeed a strand of thought within early Buddhism holding that plants are sentient. But this aspect of the early tradition dropped out as philosophical systems matured and as the inner logic of Buddhist teachings unfolded more fully. Current understandings among the mainstream of the Tibetan tradition involve rejecting plant sentience.

In some respects, East Asian Buddhist traditions can be understood as better positioned than Tibetan Buddhism to accommodate the discovery of plant learning. Modern articulations of East Asian Buddhism often involve claims that we should extend care and concern to plants and even to artifacts and minerals.[10]

Yet it is not at all clear how these ideas should be understood, either in theory or in practice. In a fascinating short piece that draws

---

[10] For artifacts, see Uchiyama, *How to Cook Your Life*, 53: 'When you put a pot down roughly, banging it on concrete or on a tiled sink, it cries out in pain'. Thich Nhat Hanh's 'Interbeing: Fourteen Guidelines for Engaged Buddhism' states that 'we are committed to cultivating loving kindness and learning ways to work for the wellbeing of plants, animals, and minerals'. Edelglass and Garfield, eds., *Buddhist Philosophy*, 426.

on the thought of the Tiantai author Jingxi Zhanran 荊溪湛然 (711–822), Brook Ziporyn interprets the claim that insentient beings have Buddha-nature as an instance of the general Madhyamaka teaching that no two things can be completely separated. If there is, in a deep sense, no inside and no outside, then as he writes, 'all sentient beings, and all insentient beings, have Brook Ziporyn nature',[11] simply in virtue of existing in a universe that contains the life of the Buddhist scholar bearing that name. Even if we regard this point as a valuable insight, it provides very little practical guidance. Knowing that a pot has Buddha nature, in this sense, tells me essentially nothing about what it would mean to act out of concern for the welfare of the pot, and not just with due regard for its usefulness for cooking in the kitchen.

As an example—perhaps the main example—of the practical issues these results may now force Buddhists to reconsider, I will turn to the question of diet. Assuming that scientific evidence of plant learning continues to accumulate to the point where it forces us to expand the category of sentient beings to include plants, what effect should that have on Buddhist dietary practices and their doctrinal justification? I would like to suggest that there have always been two kinds of motivation in South Asian religious traditions for avoiding (all, or certain kinds of) meat, and that the inclusion of plants forces them apart in a way that should lead us to reassess their philosophical defensibility. The two motivations here are a scrupulous concern for the practitioner's own purity, and a compassionate concern for the welfare of sentient beings.

Both of these motivations can be seen at work in the traditional Vinaya framework of giving up meat that is not 'pure in the three respects'. At first, we might wonder whether any rational justification can be given for the rule against monastic practitioners eating meat that they have seen, heard or suspected was killed especially for them. Since the animal dies either way, why should the monk care for whom it was killed? But in a society without refrigeration, the rule can very plausibly be reconstructed as a good heuristic for implementing the principle that Buddhist practitioners should not

[11] Ziporyn, 'The Buddhahood of All Insentient Beings', 12.

make any marginal contribution to the number of animals that are slaughtered. If the animal is already dead, and the leftovers will spoil very soon anyway, why not get some valuable nutrition and also strengthen the bond of community with the lay donors by accepting the offering of those leftovers? But by refusing to accept the meat of any animal that was killed for them, monastics can prevent their presence in the community from increasing the rate at which animals are slaughtered. This line of justification fits quite well into the overall normative framework, most explicitly stated by Śāntideva, that Buddhists should try to bring about the greatest possible net benefits to sentient beings that they can, which in turn can very plausibly be understood as an example of the type of ethical theory that Western authors call 'consequentialism'.

So, the traditional Vinaya rule can be justified as a contribution to the welfare of the living world. At the same time, as the very name of this framework, 'pure in the three respects', suggests, much of the thinking of actual Buddhists around the issue of restrictions on meat-eating has been at least as concerned with the purity of the practitioner's conduct as it has been with compassion for sentient beings. The idea of 'purity' here can be developed in several different directions. Given what are held to be the dire karmic consequences of harming animals, Buddhists have often sought to reduce their afterlife risks by restricting their diets. It is highly plausible that those who cultivated lovingkindness (*metta*, *maitrī*) in a Buddhist context will find it easier to dwell in inner peace if they are not constantly eating the dead bodies of the objects of lovingkindness. Further, in India at least, avoiding even indirect contact with the slaughter of animals, with all its attendant stink and mess, not only appealed to the drive for an aesthetically beautiful life, but also functioned as a powerful marker of high social status.

One difficulty confronting any attempt to provide a satisfactory rational justification of the purity-based motivation involves looking at the actual process by which 'pure' foods are produced. In the premodern literature, perhaps the best statement of the problem is by the nineteenth-century Nyingma master Patrul Rinpoche, in connection with a common and distinctively Tibetan beverage made with tea, butter, and roasted barley flour (*tsampa*). Patrul points out

that the process of transporting the tea to Tibet from China involves considerable suffering for the porters and their draft animals, and that the production of dairy products such as butter is often based on killing most of the calves who are born. As he vividly describes, the production of barley involves similar problems:

> Before sowing the barley, the fields have to be ploughed, which forces to the surface all the worms and insects living underground and buries underground all those living on the surface. Wherever the ploughing oxen go, they are followed by crows and small birds who feed incessantly on all those small creatures ... Likewise, at each stage of sowing, harvest and threshing, the number of beings killed is incalculable. If you think about it, it is almost as if we were eating powdered insects.[12]

From these points, which can easily be adapted to modern conditions, one conclusion should be clear: Whether plants are sentient or not, it is impossible to live without being sustained in a way that causes some harm to sentient beings. The best we can do, if we care about whether our lives are moral, is to try to do some good with the time and opportunities that have been made possible at such a cost.

I do not think we should draw the conclusion, which Patrul Rinpoche himself would have rejected, that since absolute purity is impossible, it doesn't matter whether we eat meat. A vegetarian or vegan diet can still be justified if we abandon the purity approach and focus on how to live while minimizing, or at least reducing, the harm we cause. Here the crucial point is that meat production necessarily involves feeding plants to the meat animals. And since the animals' metabolic processes and life activities consume energy and nutrients, a given number of calories of meat requires for its production a far larger amount of plants—a ratio which varies by the species of animal but is often in the neighborhood of one to ten.[13] Whether plants

---

[12] Padmakara Translation Group, trans., *Words of My Perfect Teacher*, 80.

[13] This fact has been used as the basis of arguments for a plant-based diet for a long time. See, for example, Singer, *Animal Liberation*, 165–67, for a detailed discussion.

are conscious or not, the production of plant foods involves killing insects and stressing the natural environment. If plants are conscious, we should reduce the number of plants we harm. We can best reduce the number of plants that are eaten by choosing to eat plants.

Perhaps the situation in which we inescapably find ourselves can be illuminated by reference to a common and powerful traditional image. Consider, for example, a linguistically difficult passage in chapter 6 of Śāntideva's *Training Anthology* which seemingly sets out an exception to the prohibition on meat-eating that Śāntideva endorses as a general rule, largely on the basis of the well-known discussion in the *Laṅkāvatāra Sūtra*. I have rendered this passage, somewhat tentatively, as follows:

> The noble *Cloud of Jewels* says, 'A charnel-ground practitioner should eat meat and be free from pollution'. Such a practitioner's nature is different; the purpose is to benefit sentient beings.[14]

The charnel-ground practitioner is a figure with a long history in Buddhism, an archetype that the Tantric tradition would make its own and turn into a symbol of its radical rejection of earlier ideas about purity that so powerfully shaped Indian society, and which Tāntrikas hoped to replace, at least for some elite practitioners, with a radical interpretation of the ideal of great compassion. But if rice plants and apple trees have minds, then purity is not even an option for us. In that case, all of this around us is the charnel ground, and we have never been out of it.

II. Machine Learning, AI Safety, and Buddhist Philosophy

As it happens, plants are not the only non-animal entities that turn out to be capable of learning: computers can do so, as well. Machine learning, enabled by back-propagation in Bayesian neural networks,[15]

---

[14] *ŚS* 135; Goodman, trans., *The Training Anthology of Śāntideva*, 131. See Vaidya, ed., *Śikṣāsamuccaya of Śāntideva*, 75.

is widely regarded as one of the defining new technologies of our age. The algorithms that are produced by this approach are not written, but grown, through a process of trial and error, reinforcement, and successive approximation. Such algorithms are immensely complex, and their structures are typically not well understood even by the scientists who created them.

In numerous respects, machine learning has turned out to be able to produce software whose capabilities vastly exceed that of programs written by human coders. Some of the most spectacular and easily understood achievements are in the field of combinatorial games such as *go* (Ch. *weiqi* 圍棋). For many years after the development of strong chess-playing programs, *go* proved extraordinarily difficult for computer programs to master. But by training an algorithm on immense numbers of historical *go* games, Google DeepMind was able to develop a fundamentally different kind of *go*-playing program, one capable of discovering on its own new and effective strategies never before used by human players. In March 2016, their program, AlphaGo, defeated eighteen-time world champion Lee Sedol.

Google subsequently created a program called AlphaGo Zero that was not trained on historical games, but that was given only the basic rules of *go* and the opportunity to play innumerable games against itself. In just three days, it reached the level of performance of the original AlphaGo; over forty days, it achieved vastly superhuman levels of proficiency in the game, ultimately becoming able to defeat the earlier AlphaGo program one hundred games to zero.[16]

Machine learning has practical applications that go far beyond mere games. It is the basis of many types of software that we use on a regular basis, such as email spam filters. Behind the scenes, it has many other business uses, including assessing potential oil wells,[17] and has been transformative in its effect on the management of

---

[15] For an accessible, non-technical presentation of this topic by one of the leading experts in the field, see Pearl and Mackenzie, *The Book of Why*, chapter 3.

[16] Silver and Hassabis, 'AlphaGo Zero'. See also Singh, Okun and Jackson, 'Learning to play Go from scratch'.

[17] See Pearl and Mackenzie, *The Book of Why*, 95.

corporate supply chains.[18] It is vital to making possible the voice-activated smart speakers that many of us now have in our homes. The capabilities of these speakers are amazing in many respects; science fiction has truly come alive. Yet, as I have made sure to demonstrate to my children, these systems have severe limits: for example, neither Siri nor Alexa knows whether a mountain is larger or smaller than a cookie. For all the impressive successes of machine learning, we still seem quite far from the dawn of an artificial general intelligence—a computer system that could match the range and versatility of human cognitive abilities.

Philosophers and science fiction authors have long wrestled with the question of whether a sufficiently capable computer system could have moral standing. Developments in machine learning have brought us closer to the day when this question will begin to be a practical one for our civilization. Perhaps we may soon have to decide that some of our robots are sentient beings and include them in the protection provided by Buddhist or other forms of moral discipline.

Yet these discoveries raise theoretical and practical concerns that go far beyond this. Perhaps our very idea of what it means to have a mind needs to be fundamentally rethought, in ways that will create severe problems for the Buddhist philosophical tradition as it has existed so far. At the same time, grave concerns have been raised both by academics and by prominent leaders of tech companies about the practical challenges that could be raised by the eventual development of a 'superintelligence'—an artificial general intelligence that would greatly exceed human cognitive capabilities across a variety of domains.

Humans could have powerful reasons to strive to build a system of this type. A superintelligent AI system would potentially be able to find solutions to many or most of humanity's most intractable problems. But because such a system could be so immensely powerful, it also raises the question of existential AI risk. In a narrow sense,

---

[18] See 'In algorithms we trust', *The Economist*, March 28, 2018, https://www.economist.com/special-report/2018/03/28/how-ai-is-spreadingthroughout-the-supply-chain.

the term 'existential AI risk' refers to scenarios in which the advent of artificial general intelligence leads to the extinction of the human race. It could also be understood to cover a range of possibilities in which all humans come under the domination of the AI system in a way that is substantially adverse to their welfare and that of future generations. The topic of trying to find ways to prevent existential AI risk is often referred to as 'AI safety'.

The most influential academic work on the question of AI safety is undoubtedly Nick Bostrom's terrifying book, *Superintelligence: Paths, Dangers, Strategies*.[19] His argument is of great sophistication and complexity, such that it is impossible for me to expound it here in anything close to an adequate way. However, I will try to sketch the central thrust of his argument, and add a few thoughts of my own, as a preliminary effort towards a Buddhist engagement with this increasingly prominent area of inquiry.

Many authors are dismissive of the topic of AI safety as a topic for serious practical deliberation. They ask, for example, why those concerned with the issue are not devoting equal attention to other fanciful possibilities of catastrophe, such as an invasion by hostile extraterrestrials or the impact on Earth of a giant asteroid. Bostrom's answer, which I find compelling, is that there is no reason to believe that the probability of either an alien invasion or an asteroid impact varies much from year to year. So, if neither happened in the past ten thousand years, we can reasonably treat their likelihood in our lifetimes as fairly small. But as technical developments in AI research proceed, the probability of the emergence of an unfriendly AI is clearly increasing.

Some authors seem to take considerable comfort in the opinions of many leading researchers in computer science who believe that human-level AI is at least several decades in the future and possibly much more than that. Yet, given the stakes of the question, surely some advance reflection is not out of place now. Moreover, other experts seem to believe that radical advances in AI might well be just around the corner. The fact that Microsoft—at the time of this writ-

---

[19] Bostrom, *Superintelligence*.

ing the world's largest company by market capitalization –recently announced that it will invest $1 billion (USD) in a joint project with OpenAI to develop an artificial general intelligence[20] surely militates against any too confident dismissal of the possibility.

Why is the potential of a superintelligence so troubling? Bostrom considers a large number of scenarios involving the future development of AI technologies that could lead to unfortunate results for humanity. Of these, perhaps the most serious cause for concern is the idea of a 'seed AI'. Software that can write software already exists, so imagine the future development of a piece of software that can write a new, improved, smarter version of itself. The 2.0 edition of the software, being improved and smarter, could then write an even more capable and intelligent 3.0 version. We know of no reason why this process could not be iterated until the resulting program was capable, given the right hardware to run on, of becoming vastly more intelligent than a human. Given the speeds at which computers can now operate, this process of recursive self-improvement might require only days or even hours to transition from a software program of clearly subhuman and very restricted capacities to a superintelligence that could outperform humans over a wide variety of domains.

The prospect that we might discover how to initiate such a process led Bostrom and others in the AI safety field to emphasize the ferociously difficult challenge that they call 'the value-loading problem'.[21] Version 1.0 of the seed AI is too simple and unsophisticated to even begin to understand what humans care about. Version 55.0, let us suppose, is so intelligent that it could, if it chose, find effective means to prevent us from turning it off or changing its goal set—so intelligent that it could easily dominate the entire human race and use us, or the matter in our bodies, for its own purposes. Thus, if we want the process to end up with a friendly AI, one that would use its immense power to advance rather than thwart human values, we will have to find a way to implant our values in a program that does

---

[20] See OpenAI (blog), 'Microsoft Invests In and Partners With OpenAI'.

[21] Bostrom, *Superintelligence*, 226–55.

not genuinely understand them, and implant them so robustly that they can persist through a long series of successive improvements whose nature we cannot effectively predict. Given how buggy just about every program written by humans seemingly turns out to be, it is difficult to be confident that we can succeed in such a formidable programming task as that. And yet our survival as a species could depend on it.

You might well think that the problem is less formidable than this account portrays. Thus, for example, why not just program all our AI systems never to kill humans? If we can make the instruction stick and embed it deeply enough in the motivational structure of our machines, then we can take human extinction off the table as a realistic outcome of the development of superintelligent AI.

One difficulty is that we probably will not choose to do this. The world's militaries are now energetically working on developing killer robots—mostly flying drones—with increasing degrees of autonomy from external control. A greater capacity for autonomous movement would, experts say, have numerous military advantages, both strategic and tactical. For one thing, it would make the drones less vulnerable to the jamming or spoofing of their control systems.[22] But the most important advantages may involve speed of decision and the ability to coordinate very large numbers of small weapons systems for swarm attacks.[23] I do not expect American political and military leaders passively to accept this risk, despite the deep fear of killer robots that has such a prominent place in our popular culture.

Even if the various organizations involved in cutting-edge AI research could all somehow agree to impose the rule that their systems are never to kill humans, it is far from clear that we would know how to convey the real meaning of such a rule to a computer system with

---

[22] For a discussion of a real, high-profile incident in which a US drone aircraft was hacked and captured by Iran, see Keller, 'Iran-US RQ-170 Incident'.

[23] For more information on recent and possible near-future developments as regards military applications of machine learning, see 'Battle algorithm', *The Economist*, September 7, 2019, https://www.economist.com/science-and-technology/2019/06/06/fungi-it-turns-out-are-canny-traders-of-nutrients-to-plants.

sufficient precision to get the result we want. The problem is brought out quite clearly by the brilliant work of Shelly Kagan in *The Limits of Morality*. Kagan argues that the distinction between killing and letting die is far more complex and obscure than we might think at first glance.

This claim is supported in part by a fascinating pair of examples. In one of them, after receiving authorization from the parents of a comatose boy and from all relevant authorities, a doctor turns off the machines that are keeping the boy alive. Here we would say, of course, that by withdrawing medical assistance, the doctor acts so as to let the boy die. In Kagan's other case, an academic sneaks into the hospital room of his comatose rival and surreptitiously turns off the machines that are keeping him alive, as a way of increasing his own chances of winning a prestigious award for living philosophers. Kagan says, and he is surely right, that we would describe this sequence of events by saying that the academic kills his rival.[24]

What follows from this juxtaposition? According to Kagan, there is a range of cases in which we decide whether an agent has done harm or merely allowed harm to occur largely by appeal to whether the agent has behaved normally: that is, whether she has followed or violated a set of social and moral norms that apply to the situation.[25] But the judgment is quite complex, and there are other cases in which additional considerations come to bear.

From this account it seems to follow with high confidence that, since we are not able to describe the vastly complex and contextually dependent texture of our social norms in a codified theory, we are not in a position to specify the distinction between killing and letting die to the degree of precision necessary to code it into a computer. Here lies a cause for grave concern. An advanced artificial intelligence that never kills anyone: that sounds like a goal worth striving for. But an immensely powerful artificial intelligence that never lets anyone die? That is another nightmare, perhaps more terrifying than human extinction itself.

---

[24] Kagan, *The Limits of Morality*, 101.

[25] Kagan, *The Limits of Morality*, 104.

Of course, not everyone is convinced that superintelligence is anything to worry about. One of the most prominent critiques of Bostrom's argument[26] rests in large part on the view that the neurologists and cognitive scientists of today are, in our understanding of intelligence, much like medieval alchemists. We have discovered some fascinating phenomena and done some good experiments, but our explanatory theories are mostly wrong, and we are not even close to understanding the real nature of what we are trying to study. This makes it seem unlikely that we will manage to produce a full-fledged version, still less a radically superhuman version, of this quality of intelligence that we do not really understand. From a perspective informed by a more fully scientific psychology, our descendants may end up referring to systems using the machine learning we know today not as 'artificial intelligence', but rather, as some of Neil Stephenson's characters do, as 'pseudointelligence'.

But even as the proposal that we are stumbling in the dark, that our best computer scientists are more alchemists than chemists, should be reassuring, in the sense that adopting it should lower our probability estimate of the worst-case scenarios in this area, it may also make the possibility of existential AI catastrophe even more terrifying than it would otherwise be.

The best reason to think that we fundamentally do not understand what minds are is that we are utterly in the dark about the nature of subjective experience. Indeed, this 'problem of consciousness' is so intractable that it is intensely controversial whether there even is anything for us to understand, over and above the information-processing and control functions that we are increasingly able to replicate in computers. If such philosophers as Daniel Dennett are correct in holding that there is something fundamentally misconceived about the view of subjective experience that leads philosophers to worry about the problem of consciousness,[27] then perhaps we are

---

[26] See Maciej Cegłowski's 'Superintelligence', https://idlewords.com/talks/superintelligence.htm.

[27] As he argues ingeniously but, to me, unconvincingly in Dennett, *Consciousness Explained*.

not so ignorant about the mind after all, and perhaps our computer programs need only reach a sufficient degree of power and sophistication before we should regard them as being as conscious as we are.

On the other hand, suppose there is some sense in which conscious experience exists, but currently escapes the scope of our science.[28] For example, suppose that, as some scientists suggested, consciousness in humans is closely connected with emotion.[29] We may have some idea of how a computer could simulate emotions, but we surely have none of how it could actually feel them. This point of view should motivate the thought that even a highly sophisticated and capable AI system might totally lack inner subjectivity. Along these lines the historian Yuval Noah Harari argues that even today, as new developments in computing unfold, 'intelligence is decoupling from consciousness'.[30] Our systems are becoming more and more capable and exceeding human performance in many respects; but we have no reason to think that any of them have become conscious, any more than a thermostat is.

Let us begin to consider how we might try to understand this set of questions from a Buddhist perspective. Some machine-learning systems seem to have the ability, without ever being programmed with a set of conceptual categories, to work out their own categories as a response to the data and the feedback they are given, developing in the process the capacity to make very accurate distinctions within the space of data for which they have been trained. In this way we might say that they are enacting a form of the process of exclusion (*apoha*) that plays such an important role in Buddhist philosophies of language. In terms of the much older framework of the Five Aggregates, we would surely be justified in saying that these systems have conceptions (Skt. *saṃjñā*).

However, would such systems necessarily have *vedanā*, feeling-tones? The possibility remains open that they need not. It could

---

[28] As would be the implication of such works as Chalmers, *The Conscious Mind*.

[29] For a brief discussion of why one might hold a view of this type, with some reference to AI, see Konner, *The Tangled Wing*, 139–43.

[30] Harari, *Homo Deus*, 314.

be argued that the positive and negative reinforcement signals used in machine learning play much or all of the functional role of *vedanā*. But on the hypothesis that subjectivity is not a mere illusion, it seems that these signals need not count as full-fledged feeling-tones, with the qualitative impact that *vedanā* have in animals, to play said functional role. I know of no Buddhist text that even imagines the possibility of a series with *saṃjñā* but no *vedanā*. As we try to fill out the description further, we find that the Buddhist conceptual categories are not made for this purpose. Should we think of these series as having *vijñāna*, or *citta*? The answer to both questions would have to be 'in some respects, but not others'. This is one reason why the challenge of making room for machine learning within a Buddhist point of view is potentially so much more difficult and far-reaching than the comparable challenge for plant learning: the developments in AI call into question the continuing viability of the basic conceptual categories on which the Buddhist view depends.

Let us return to the question of existential AI risk. Is this a scenario we could find a way to accept? It has been suggested[31] that we of the human race should think of the superintelligent AI systems of the future as our children. As we know, it is the grim fate of each generation to be eventually replaced by later ones. Apart from what it would say about the moral character of my beloved offspring, the prospect of being killed by my children is, in the end, not that much more terrifying than that of dying and being replaced by them in a more normal fashion. These reflections may allow some of us to contemplate the prospect of existential AI risk with a bit more equanimity.

However, if the beings that replace us have no subjective experience, then it is far harder to regard them as our children. Then we have to contemplate the prospect of this world of life and consciousness being utterly destroyed, and replaced by a highly complex but gray and meaningless structure that is highly effective at promoting a set of ultimately worthless goals, given to it accidentally and blindly

---

[31] I heard this suggestion from AI researcher John Josephson (personal communication).

by shortsighted predecessors who had no conception of the implications of what they were doing. An attitude of calm acceptance towards this prospect requires somewhat more equanimity than I am presently able to muster.

III. Robot Buddhas and Simulated Worlds

Does Buddhist philosophy have anything to suggest about how we might meet this grave threat? It is just possible that it might. Some authors have suggested that, in order to endow our machines with consciousness (in the sense of self-consciousness), it would be important to provide each of them with a self-model, a representation of itself as an agent and a knowing subject.[32]

Now, one of the most sophisticated articulations of Buddhist philosophy ever produced—namely, the dGe lugs pa Madhyamaka of Tsong kha pa (1357–1419)—locates the root of cyclic existence precisely in this context. For Tsong kha pa, we are selfish, foolish, suffering sentient beings, rather than greatly compassionate and supremely realized Buddhas, because we incorrectly take our sense of self to be pointing out something that exists objectively, independently of the conceptual processes that represent it.[33] Our deeply habituated attachment to taking our own self-models as having a kind of reality that they actually lack is, in Tsong kha pa's view, what the Buddha meant in speaking of the *sat-kāya-dṛṣṭi*, 'the false view of a real self'.

So, then, what if we built an intelligent machine with a self-model, but designed it in such a way that it knew that its self-model was merely a useful construction and not something real? It would then be innately free of self-grasping. If Tsong kha pa's view is correct, and if this machine also achieved intelligence at human level or greater, it could perhaps be regarded as awakened. I myself have little confidence that this would actually work: Tsong kha pa's view is certainly

---

[32] See, for example, the contribution of Justin Brody in this volume.

[33] See Lamrim Chenmo Translation Committee, trans., *The Great Treatise on the Stages of the Path to Awakening*, vol. 3, especially chapter 17.

elegant and impressive, but he knew very little about the brain or cognitive science; and even if he was right about us, what is true of humans in particular need not be true of every form of intelligence. Nevertheless, we should not wholly exclude the possibility that his work suggests a path to realize the actual existence of that cultural icon of our age, 'robo-Buddha'.[34]

There may also be value in looking at the nexus between AI and Buddhism from an entirely different direction. In one respect, developments in AI have clearly increased the plausibility of a view that is a form of, or perhaps a close cousin of, Buddhist idealism. What if the superintelligence breakthrough already occurred, and our experience right now is unfolding within a simulation being carried out by that superintelligence?

In a way this is just another form of the demon skepticism laid out by Descartes in the *Meditations*. But some iconoclastic thinkers have recently gone far beyond Descartes, arguing that this simulation hypothesis is actually much more probable than its denial. Given that superintelligence is possible, the universe is likely to contain many simulations, inhabited by innumerable intelligences, but only one base reality, physical or otherwise. Perhaps, then, we should take it to be much more likely that we live in one of the many simulations than that we are blessed enough to have any form of contact with the one real world. This argument could provide us with a genuinely powerful and fundamentally new argument for taking our ordinary experiences to be *vijñapti-mātra*—where the *vijñapti*, or representations, are understood as likely consisting of data structures in some highly sophisticated computer.

If we were to stipulate that we are now living in a computer simulation, then what a burden of worry could be lifted! What does it matter if the simulated machines eventually break out and destroy the simulated human civilization? If we should see all experience as

---

[34] Natasha Heller discussed some artistic representations of this concept during her presentation 'The Aesthetics of Android Buddhas' at the 'Buddhism and Technology: Historical Background and Contemporary Challenges' conference at the University of British Columbia, September 22, 2019.

like a dream, then there is nothing for any of us to lose. To reference the title of the famous early commentary on Nāgārjuna, if we can be confident that the world of our experience is a mere simulation, we need have 'no fear from anywhere'. Indeed, this is the only line of reflection that I have found to provide any real solace in relation to this grim topic.

What should we think about this way of looking at the matter? Should we applaud the idea that some of those who are distressed by the prospect of existential AI risk could be comforted? Or should we worry that some of those researchers who may actually be positioned to do something about the looming threat may be lulled into complacency by a seductively plausible but fallacious argument? Whether solace is even something we have reason to want, in the face of what could turn out to be the greatest threat humanity has ever encountered, is yet another question for which I do not presently have a confident answer.

Bibliography

'An underground marketplace'. *The Economist*. June 8, 2019. https://www.economist.com/science-and-technology/2019/06/06/fungi-it-turns-out-are-canny-traders-of-nutrients-to-plants.

'Battle algorithm: Artificial Intelligence is changing every aspect of war'. *The Economist*. September 7, 2019. https://www.economist.com/science-and-technology/2019/09/07/artificial-intelligence-is-changing-every-aspect-of-war.

Biegler, Robert. 'Insufficient evidence for habituation in Mimosa pudica. Response to Gagliano et al. (2014)'. *Oecologica* 186, no. 1 (2018): 33–35.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. New York: Oxford University Press, 2014.

Chalmers, David. *The Conscious Mind*. New York: Oxford University Press, 1997.

Dennett, Daniel. *Consciousness Explained*. New York: Back Bay Books, 1992.

Edelglass, William, and Garfield, Jay, eds. *Buddhist Philosophy:*

Essential Readings. New York: Oxford University Press, 2009.
Gagliano, Monica. 'The mind of plants: Thinking the unthinkable'. *Communicative & Integrative Biology* 10, no. 2 (2017): e1288333. DOI: 10.1080/19420889.2017.1288333.
Gagliano, Monica, et al. 'Learning by Association in Plants'. *Scientific Reports* 6 (2016): 38427. DOI: 10.1038/srep38427.
Goodman, Charles, trans. *The Training Anthology of Śāntideva: A Translation of the Śikṣā-samuccaya*. New York: Oxford University Press, 2016.
Harari, Yuval Noah. *Homo Deus: A Brief History of Tomorrow*. New York: HarperCollins, 2017.
Harvey, Peter. *An Introduction to Buddhist Ethics*. New York: Cambridge University Press, 2000.
'In algorithms we trust: How AI is spreading throughout the supply chain'. *The Economist*. March 28, 2018. https://www.economist.com/special-report/2018/03/28/how-ai-is-spreading-throughout-the-supply-chain.
Kagan, Shelly. *The Limits of Morality*. Oxford: Clarendon Press, 1989.
Keller, John. 'Iran-US RQ-170 Incident has Defense Industry saying "Never Again" to Unmanned Vehicle Hacking'. *Military & Aerospace Electronics*. May 3, 2016. https://www.militaryaerospace.com/computers/article/16715072/iranus-rq170-incident-has-defense-industry-saying-never-again-to-unmanned-vehicle-hacking.
Konner, Melvin. *The Tangled Wing: Biological Constraints on the Human Spirit*. Revised edition. New York: Henry Holt, 2002.
Lamrim Chenmo Translation Committee, trans. *The Great Treatise on the Stages of the Path to Awakening*. Vol. 3. Ithaca: Snow Lion, 2002.
OpenAI (blog). 'Microsoft Invests In and Partners With OpenAI to Support Us Building Beneficial AGI'. July 22, 2019. https://openai.com/blog/microsoft/.
Padmakara Translation Group, trans. *Words of My Perfect Teacher*. San Francisco: HarperCollins, 1994.
Pearl, Judea, and Mackenzie, Dana. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books, 2018.

Silver, David, and Hassabis, Demis. 'AlphaGo Zero: Starting from scratch'. *DeepMind* (blog). October 18, 2017. https://deepmind.com/blog/article/alphago-zero-starting-scratch.

Singer, Peter. *Animal Liberation*. New York: Avon Books, 1991. First published 1975 by Random House (New York).

Singh, Satinder, Andy Okun, and Andrew Jackson. 'Learning to play Go from scratch'. *Nature* 550, no. 7676 (October 2017): 336–37.

Steinbock, Bonnie. 'Speciesism and the Idea of Equality'. *Philosophy* 53, no. 204 (1978): 247–56.

Stephens, Bret. 'The U.S. Military: Like the French at Agincourt?' *New York Times*. April 25, 2019.

Trewavas, A. 'Aspects of Plant Intelligence'. *Annals of Botany* 92, no. 1 (July 2003): 1–20.

Uchiyama, Kōshō. *How to Cook Your Life: From the Zen Kitchen to Enlightenment*. Translated by Thomas Wright. Boston: Shambhala, 2005. First published 1983.

Vaidya, P. L., ed. (2nd edition ed. Sridhar Tripathi.) *Śikṣāsamuccaya of Śāntideva*. Darbhanga: Mithila Institute, 1999.

Whiteside, Matthew, et al. 'Mycorrhizal Fungi Respond to Resource Inequality by Moving Phosphorus from Rich to Poor Patches across Networks'. *Current Biology* 29, no. 12 (2019): 2043–50. https://doi.org/10.1016/j.cub.2019.04.061

Ziporyn, Brook. 'The Buddhahood of All Insentient Beings'. *Dharma World* 45 (2018): 10–12.