A Buddhist Contribution to Artificial Intelligence?

DOUGLAS DUCKWORTH
Temple University
duckworth@temple.edu

Abstract: Significant questions confront Buddhist traditions in the wake of emergent technologies: can the human body be configured in a certain way, such that it reveals a new world or environment to inhabit beyond optimized self-preservation or survival? Can we manipulate our bodies with technologies—inhibited (or enhanced) by a chemical, a trauma, a contemplative technique, or an implant—such that we are reoriented to a transformed and liberating understanding of the nature of the world and our being in it? As new technologies enhance certain domains of cognitive performance by modelling and extending the structure and capacities of cognition, Buddhism, with a theory of mind and mental development in the absence of an independent essence, owner, or agent like a self, can potentially be a valuable resource. Buddhism provides a useful theoretical foundation to articulate not only the potentials for engineering intelligence, but also by identifying problems in this project.

Keywords: Artificial Intelligence (AI), predispositions, Buddhism, karma, cognitive science

DOI: : https://dx.doi.org/10.15239/hijbs.03.02.02

Introduction

In this paper, I will articulate a Buddhist perspective on the potentials and limits of Artificial Intelligence (AI). Over millennia, Buddhists developed resources for understanding the elusive nature of mind. Significant questions confront Buddhist traditions in the wake of emergent technologies: can the human body be configured in a certain way, such that it reveals a new world or environment to inhabit beyond optimized self-preservation or survival? Can we manipulate our bodies with technologies—inhibited (or enhanced) by a chemical, a trauma, a contemplative technique, or an implant—such that we are reoriented to a transformed and liberating understanding of the nature of the world and our being in it?

As new technologies enhance certain domains of cognitive performance by modeling and extending the structure and capacities of cognition, Buddhism, with a theory of mind and mental development in the absence of an independent essence, owner, or agent like a self, can potentially be a valuable resource. Buddhism provides a useful theoretical foundation to articulate not only the potentials for engineering intelligence, but also by identifying problems in this project.

4E Cognition

We can find resources for thinking about these challenges with the notion of '4E cognition', a framework developed from cognitive science: cognition is *embedded*, *extended*, *embodied*, and *enactive*. For instance, we clearly find direct links to *embedded cognition*, the first 'E', and the coupling of mind and world, or organism and environment, in the ways that Buddhist texts in particular have framed the constitution of the world (*loka*), a world in interrelations and the coupling of organs (*indriya*) and objects (*viṣaya*) rather than in isolated, discrete entities. In particular, the world in Yogācāra Buddhism is not observed through the peephole of a 'Cartesian Theatre', or in a sort of 'spectatorial epistemology' secured from a standpoint outside the framework. Rather, in Yogācāra, we find a rich model of cognition deeply embedded within the structure of the world.

As we see in results from cognitive psychology and Experimental Philosophy (X-Phi), the embedded, situated nature of cognition matters. Context matters. For instance, people have been shown to evaluate other people differently when holding a cold cup of coffee in contrast to a warm one,[1] and people make more severe moral judgments when in dirty rooms strewn with dirty pizza boxes and given chewed pencils to answer questions.[2] Clearly, a rich understanding of the contexts of cognition and judgment is important to take into account to develop a robust model of self-understanding in new situations afforded by the shifting terrain of the modern world.

Since humans survive by filtering information from a sensory manifold—discrete acts of knowledge are directed (intentional) and *attention is paid*. Thus, when the processes of filtering and directing takes place, what are the goals (implicit or otherwise) presumed by the decisions to notice one thing at the expense of another? Since data only comes to be *information* when decisions are made (as to which differences make a difference), when robots with high-tech cameras and sound-receptors register waves of light, sound, etc. beyond the spectrum of human sense faculties, what is it, or who is it, that makes the decision about which data counts as information? These questions are important for addressing the scope and foundations of intelligence, engineered or otherwise.

As with embedded cognition, the second 'E', extended cognition, also conveys the way that the mind or mental processes, beyond the brain, shape and inform the worlds we inhabit. In its strong form, extended cognition undermines any kind of hard and fast distinction between mind and matter, as we see in the subject-object nonduality in Yogācāra Buddhism. The fact that the duality presumed in a strong mind-matter distinction collapses bolsters the case for a strong potential of AI, for when cognition is not necessarily unique to human organisms, it is a quick step to affirming consciousness everywhere (in a gorilla, a dog, a fish,

---

[1] Williams and Bargh, 'Experiencing Physical Warmth Promotes Interpersonal Warmth', 606–7.

[2] Schnall, et al., 'Disgust as Embodied Moral Judgment', 1096–109.

a flower, a cell, a planet, a robot, or a thermometer). After all, the mind is there wherever we turn our attention. Yet we can blow up the brain to billionfold levels of magnification and find no central location or anything in the brain that can be said to be 'mind'. The mind is everywhere yet nowhere.

Learning to use a prosthetic leg, for instance, is a classic example that shows the way that cognitive maps extend beyond the boundaries of a skin-encapsulated ego. The line between mind and matter also blurs as we come to extend our body schema to include technologies like an artificial limb or a machine—a bicycle, a car, or a phone—as we continue to orient and reorient ourselves to navigate new tools and environments. This is the case with body-mind enhancement or impairment (when something 'artificial' is incorporated into the body-schema, like eye-glasses, limbs, chemicals, or neural implants) that shape hybrid systems of human-machine cyborgs.

The third 'E' is embodied cognition. We are confronted viscerally with the nature of embodied cognition when something stops working. Amputees, for instance, continue to feel their severed 'ghost limbs', while sufferers of anorexia see their emaciated bodies as fat. The embodied nature of cognition is closely tied to its embedded and extended nature, too, as we hardly notice a doorknob until it stops working, or until it is installed in a frame and put on display in a contemporary art museum, calling our attention to see it in a new way.

The last of the four 'Es', enactive cognition, challenges a representational theory of meaning. For instance, a machine has limited power when designed to represent, not interact, with a flat world. A computer that is designed simply to make numeric computations will have predictive successes exceeding that of a human, but like a severely autistic child, will not necessarily be able to adapt readily to a multifaceted world. Like a system designed solely for computing one value, like the *homo economicus*, 'the Economic Man', who limns the world solely in terms of self-interested, monetary values, the unhinged reductionism of data-driven machines can wreak havoc on the health of society and the world, even if machines are better than humans at calculations and games like chess and Go. What constitutes 'intelligence' remains a question here, and how

intelligence comes to be differentiated from 'life' or 'mind' (e.g., is an eagle more 'intelligent' since it can fly and see farther than a human?). Neuroscience may develop beyond mechanistic models of mind towards increased complexity, and as the technologies of machines change, so do conceptions of mind. The rich complexity of cognition needs to be taken into account to more fully understand and model intelligence.

Consider the 'trolley problem', for instance, to draw attention to the ethical dilemmas of machine design. One version of the 'problem' goes something like this: when people are asked if, given the opportunity, they would sacrifice one person in order to save five, of course almost everyone says they would do it. Then there is a richer description, whereby one is confronted with a dilemma: whether or not to push a fat bystander onto the tracks to stop a train that is about to roll over and kill five people tied to the tracks. Would you do it? Here, many people change their answers to 'no'.

On one level, the trolley problem evokes nothing else than the dissonance between theory and practice. The 'trolley problem' is a problem of abstract theory divorced from enaction in real situations in the world. In other words, it is a problem of simulation. In living situations in the world, things are not based on binary systems of either/or, black or white, right and wrong. Also, there are no guarantees. There is no guarantee that when you push a fat guy he will fall on the train tracks where you want him to; and there is no guarantee whether or not he will pull you down with him if you push him. Further, there is no guarantee whether or not you will go to jail after you push that man, or that after he dies, the train will stop before anyone is killed. For these reasons, these kinds of simulations reveal the dissonance between abstract theorization and the lived world. Yet what we can gain from data gathered from tests like these might allow us to model collective human experience in ways that decenter the entrenched habits of ego-driven decisions.

For instance, we can see how this simulation translates into the design of self-driving cars: would you choose a car that would kill your spouse rather than run over a stranger on the road? Would you want your car to sacrifice the owner rather than kill a seventy-year-old grandmother? Her grandchildren? Answers to these kinds of ques-

tions (surveyed in a crowd-sourced project at MIT)[3] may come to dictate decisions in AI design. These questions also are rooted in ancient debates over human values, and Buddhism has resources for not only bringing these kinds of ethical dilemmas into focus, but for training individuals to deal with them beyond the primitive habits of narrow, self-centered interest. Thus, we can find in Buddhism something more than just abstract theory, and something that does not simply confirm a 'natural' reaction—to act out an instinct for self-preservation. Rather, Buddhism offers a unique set of theories and practices for re-orienting our habitual cognitive and behavioral responses to the world.

Language, Hardware, and Software

A Buddhist view highlights how our decisions are primed by predispositions we inherit (our karma, our genes) that are buried under the surface, in our gene pools and habits inherited from a collective past of our human and non-human ancestors. One of the problems and promises of machine design is that machines, as cultural products, reflect the psyche and goals of their creators—our machines are an extension of ourselves and an expression of human values. With the development of artificial intelligence lies the hope to continue an evolutionary path to actualize the potential intelligence beyond that of the current human condition—the *post-* or *transhuman*. Yet a potential danger lurks, like in Mary Shelley's *Frankenstein*, when we make machines in our own image (consciously or not), and our habitual tendencies are embedded subconsciously, or unconsciously, in our creations. That is, we can easily perpetuate the structures of violence and oppression that dwell below the access consciousness of our conscious minds—in our bodies. This is another place where a potential Buddhist contribution to machine design and development may be found.

Machines replicate aspects of the psyche (intentions and interests)

---

[3] See Moral Machine (website).

of the designer and the societies and institutions that support technological design, so if engineered desires lies in the future of technological innovation, this may be Buddhism's most important contribution. As Yuval Harari concludes his bestselling book, *Sapiens*: 'Since we might soon be able to engineer our desires too, perhaps the real question facing us is not, "What do we want to become?", but "What do we want to want?"'[4]

Future speculations aside, we can catch a glimpse now of the deep structure of unconscious forces at play in human experience when we consider cases of implicit bias. Implicit bias tests reveal how most of us tacitly embody negative stereotypes of particular races, genders, and sexual orientations, despite the fact that we explicitly deny these biases. A promise of technology is that—as with the case of the machine-assisted learning that can help expose these biases held beneath the surface—machines can enable us to learn by supplementing the naïve intuitions on the surface of first-personal, phenomenological accounts of the world. A danger of technology is that it continues to replicate and extend perverse implicit structures, making innate habits (such as self-centeredness) writ large.

A clear acknowledgment of pervasive distortion, and how knowledge is always *interested* and *embedded*, may be particularly relevant for artificial intelligence design. The distorted baseline of ordinary cognition is theorized in the Yogācāra Buddhist theory of predispositions (*vāsanā*). In particular, three types of predispositions developed in the *Mahāyānasaṃgraha* are relevant to consider here: (1) 'predispositions for a view of self' (*ātmadṛṣṭi-vāsanā*), (2) 'predispositions for the branch of existence' (*bhavāṅga-vāsanā*), and (3) 'predispositions for linguistic expression' (*abhilāpa-vāsanā*).[5] The *predispositions for a view of self* names the structure by which self and other come to be bifurcated. It is said to be the cause for the illusion of a unified self (*satkāyadṛṣṭi*). The *predispositions for the branch of existence* labels the causal process that leads to embodiment within a particular realm of existence. This category of predisposition offers

---

[4] Harari, *Sapiens*, 464.

[5] Asaṅga, *Mahāyānasaṃgraha* I.58.

an account of the integral relationship between mind and world in a kind of organism-environment coupling.

In the function of this second class of predispositions we find a process where the environment changes with its inhabitants, in contrast to a Darwinian idea of adaptation within a static, natural (external) habitat. In this Yogācāra theory, cognition and habitat co-evolve (or co-*devolve*) in a process of 'cognitive niche' construction.[6] These first two types of predispositions constitute important elements to consider for understanding the ways that individual and intersubjective worlds come into shape.

The third type of predisposition, the *predispositions for linguistic expression*, labels the innate naming and labeling tendency itself; it is nothing less than the language instinct. Whereas the *predispositions for a view of self* is akin to a generic instinct that drives self-preservation, and the *predispositions for the branch of existence* is explicitly linked to the virtuous and unvirtuous actions that guide the perpetuation of species-specific forms of life, the *predispositions for linguistic expression* describes what is most developed in the human species, language. This kind of predisposition thereby differentiates what we might call processes of cultural acquisition and change from innate adaptive capacities that are more biologically hardwired. Biological change is driven largely (if not exclusively) by forces beyond conscious control. The capacity for language, too, is one that is unconscious, but the higher-order processes of thought enabled by language add another dimension to the complex of forces guiding individual, social, and technological change.

Language is a superstructure that acts as a subconscious repository exerting influence on its constituents; it contains 'big data' that can be mined for subliminal information and unconscious forces, while functioning beyond what any particular individual acting in isolation is capable of doing. Also, language gives shape to narratives of self-identity while functioning in the absence of any real basis of designation for a single actor or agent behind the scenes of its stories.

---

[6] On the 'cognitive niche', see Pinker, 'Language as an Adaptation to the Cognitive Niche'.

Language also provides a tool with which this process of narration can be tracked and guided, as well as the means by which a story can be told to be just that: a story. Interventions in the narratives that shape purportedly innocent notions of essence and identity—in other words, *critique*—may be the most relevant place for Buddhist insights to weigh in on the inputs and outputs of AI.

There is also clearly an advantage with a machine's ability to intervene with our naïve intuitions. We can see this in the case of perceptual illusions, where first-personal reports conflict with third-personal data, even when we are primed to know that the illusion is an illusion. One of the deliverable promises of science is to innovate technologies to overcome naïve intuitions (or at least lead us to recognize our susceptibility to this kind of unconscious error). By incorporating the perspective of a disinterested, third-personal gaze toward the world, we can cultivate a critical and broader understanding of the world and our place in it.

Buddhist theories of no-self in particular can help us think about how we can integrate first- and third-personal perspectives. The ideal of complementary rather than conflicting relations among perspectives, I contend, is a preferred alternative to replicating the cycles of oppression and violence that stem too often from a model of subjects competing against objects, stemming from a bifurcation of (1) the first personal perspective, enabling the tribalism that is the 'us *vs*. them', 'I and it' mentality, and (2) the cold and abstract third-personal gaze that sees others and the earth exclusively in terms of objects, impersonal data, and commodities (e.g., natural resources).

Conclusion

A Buddhist contribution to AI, I contend, lies primarily in its powerful critique, and by charting the ways that conceptual and linguistic static disrupt a proper understanding of the way things are, particularly when one node of an intertwined network (e.g., a self or other discrete entity) is reified and privileged at the expense of another. Unlike the modernist notion of *raw data*, which is presumed to be disinterested and uninterpreted, what distinguishes informa-

tion (and knowledge) from data is that it is tied to interests. The fact that knowledge is always interested—and embedded, extended embodied, and enactive—is acknowledged in a Buddhist account, where the notion of intelligence is not premised upon an ideal of disembodied and disinterested data any more than its evolution is marked by a course of self-preservation. Rather, an important function of intelligence in Buddhism is expressed through critique—one that unwinds or breaks a vicious cycle by undermining its driving force.

The places where a Buddhist perspective can contribute to understanding both the promises and dangers we face in the development of new technologies like AI can thus be found primarily in its articulation not only of posthuman potentials, but in human limits. In particular, Buddhist traditions have many resources to draw from to forcefully critique misguided notions of intelligence—technologically-enhanced or not—specifically, those that falsely presume that human knowledge and mechanized data innocently and accurately represent the world from a fundamentally neutral, natural, and disinterested stance.

Bibliography

Primary Sources

Asaṅga. *Compendium of Mahāyāna* (*Mahāyānasaṃgraha*, *theg pa chen po'i bsdus pa*). D. no. 4048.

Secondary Sources

Harari, Yuval Noah. *Sapiens: A Brief History of Humankind*. New York: HarperCollins, 2015.

Moral Machine (website). Scalable Cooperation, MIT Media Lab. Accessed November 11, 2020. https://www.moralmachine.net/.

Pinker, Steven. 'Language as an Adaptation to the Cognitive Niche'. *Studies in the Evolution of Language* 3 (2003): 16–37.

Schnall, Simone et al. 'Disgust as Embodied Moral Judgment'.

Personality & Social Psychology Bulletin 34, no. 8 (2008): 1096–109.

Williams, L. E., and J. A. Bargh. 'Experiencing Physical Warmth Promotes Interpersonal Warmth'. *Science* 322 (2008): 606–7.