Journal of Buddhist Ethics
ISSN 1076-9005
http://blogs.dickinson.edu/buddhistethics/
Volume 27, 2020

True Love for the Artificial? Toward the Possibility of Bodhisattva Relations with Machines

Thomas H. Doctor

Kathmandu University Centre for Buddhist Studies at Rangjung Yeshe Institute

True Love for the Artificial? Toward the Possibility of Bodhisattva Relations with Machines

Thomas H. Doctor [1]

Abstract

Given our increasing interaction with artificial intelligence and immersion in virtual reality, which epistemic and moral attitudes towards virtual beings might we think proper, relevant, and fulfilling? That is the basic question that this article wishes to raise. For the main part, it presents a descriptive analysis of our current situation, which is meant to expose features of artificial intelligence (AI) and virtual reality (VR) that seem both salient and easily aligned with central Buddhist concerns. Developed without any requirement for, or expectation of, the existence

[1] Kathmandu University Centre for Buddhist Studies at Rangjung Yeshe Institute. Email: thomas.doctor@ryi.org. Many thanks are due to my esteemed colleagues at the Center for the Study of Apparent Selves (www.csas.ai) with whom I have the good fortune of exploring the issues that this paper takes preliminary notice of. We gratefully acknowledge the generous support of the Templeton World Charity Foundation. Thank you also to the *JBE* reviewer, Rick Ripetti, for his rich and constructive feedback.

of real subjects and selves, Buddhist views and practices clearly resonate with the assumptions of unreal mind and mere appearance that are associated with AI and VR. Yet Buddhists famously also declare that the illusion-like nature of things does not negate, but in fact entails, universal care and deep meaning. I conclude by suggesting that such doctrinal claims may be tested for practical relevance in the present and emerging world of interconnectivity and illusion.

Introduction

Whatever we may want to say about the likelihood of human or superhuman artificial intelligence in the near or long-term future, our turbulent world is quite undeniably in a process of embracing and being embraced by increasingly intelligent machines. We may approve of and appreciate that development, we might reject and condemn it, or we may suspend judgment and remain agnostic—in fact, most likely, we find ourselves responding in all those ways, depending on how and why we are approached. But regardless of our attitudes and opinions, the encounter with virtual bodies and minds is pervasive. From the crude to the highly sophisticated and abstract, we tend to meet and interact with human-engendered, emergent intelligence throughout our public and private lives. Our continuous interactions by means of the pocket device that we—despite its myriad functions—typically still just call our "phone" may serve as a good illustration. The contexts and outcomes of such encounters and exchanges are arguably no less real and solid than old-fashioned inter-subjective exchange. If we simply look back at our lives during the past twenty-four hours, we may find that some of the most impactful events were embedded in an environment enabled by AI. Per-

haps I found myself online-diagnosed with a virus through an AI-powered health service; perhaps I shared my innermost hopes with a beloved person who is physically distant but fully present online; perhaps I had a so-called virtual meeting with colleagues, discovering that our work had been rewarded or that I had lost my job; perhaps I voted "virtually," or enrolled as a member of a club, and so forth.

No one wonders where the internet begins or ends, nor does it make much sense to ask "who is there" inside it. At the same time, many of us take its enabling presence for granted just as much as the air we breathe (especially if we live in places not particularly plaqued by air pollution). Arguably, we are informed and affected just as much by engaging within a visibly open-ended web of transpersonal cognitive factors as we are by face-to-face and heart-to-heart exchanges in the flesh. But the common intuition that an intelligent, live meeting implies interchange between separate individuals, or distinct selves, is undermined by our undeniable perception of fluidity and emergence. "What was I really looking at, whom was I actually talking to?"—when we are not wholely absorbed in the action, such questions tend to creep in during, perhaps, a pause in our immersive video game; or when we perceive our loved one on the tablet screen, vividly present but wholely untouchable; or when we navigate our way through rush hour traffic in a foreign city, all the while communicating with the delightfully familiar voice of our GPS. When such challenges to the common sense of self are particularly amplified we tend to prefer qualifying our lucid interlocutors as "artificial intelligence" (AI) and the general framework of the given events may then further be deemed a mere "virtual reality" (VR).

The move to qualifying certain aspects of our lives as less than fully real is arguably a natural one—how, we might say, would we make sense of dreams and hallucinations without recourse to the notion that things can be much less than they seem. Indeed, failing to make such dis-

tinctions in the proper way is considered a severe pathological condition. Yet although we may well dismiss the city we saw during our sleep as "just a dream" when we wake up in the morning, our dream can of course still affect us and we do not need to be psychoanalysts to believe that we can also learn from a dream. In other words, dreams have, of course, a reality of their own and when we dismiss them as unreal we know very well that is not the final word on the matter. The question of reality is, for all intents and purposes, a question of gradation, and gradation requires context. For example, Descartes discovered analytic geometry in a dream, emphasized that 2+2=4 whether dreaming or waking, and clearly thought analysis of the difference between dream and waking states was important epistemological and metaphysical work. Discovering a branch of mathematics while in an altered state, or being able to affirm a mathematical truism within an altered state, both call into question the general notion that anything occurring in altered states is disconnected from reality. One need not enter upon the sort of elaborate investigation into the differences between the altered state of dreaming and waking reality Descartes delved into in order to see the connection between such issues and AI and VR.

Now, in the current context of increasing interaction with artificial beings and immersion in virtual reality, which epistemic and moral attitudes towards virtual beings might we think would be proper, relevant, and fulfilling? That is the basic question I wish to raise through this paper. In raising it, I also wish to suggest that Buddhist insights and ways of life could prove to be useful resources for a relevant answer. Although this paper hence hopes to encourage constructive efforts in Buddhist philosophy, I will here not attempt to develop, in any detail, a particular Buddhist position on AI and VR. Instead, my approach will be a descriptive analysis of the current context, which is meant to expose features of the development of AI and VR that seem both salient and easily aligned

with typically Buddhist concerns. Some basic elements of such an alignment will then be extrapolated in the paper's final section.

§ 1 Living the AI Life: Feeling in and out of Touch

Intuitively, AI and VR may, like beings and environments in dreams, be considered less than fully real. Nevertheless, despite the qualifiers "artificial" and "virtual" there are also good reasons for qualifying the significance of AI and VR events as superior to and even more substantial than most standard, interhuman encounters and communications. Indeed, the skillful AI that I communicate with on my telephone is so well-informed that I tend to trust her/him/it far more than any human I might otherwise communicate with. I turn to the internet—and AI—when I'm faced with decisions that I feel require careful thought and vast, substantial knowledge. The contrast with the world of dream is striking—no matter how important we may think our dreams.

Noticing this simultaneous pull toward two opposing epistemic attitudes seems important. On the one hand, working with virtual objects and artificial beings comes with a sense of being distanced from the real, a sense of being out of touch with the natural. Yet as our actions bear evidence of, we also conclude that a relationship with precisely such objects and beings will afford us the greatest knowledge of the realm of the real and the natural. Firm knowledge of facts, we feel, is achieved through comprehensive cooperation, if not outright immersion, with artificial intelligence. If this observation of two opposing but coexistent intuitions about AI and VR is correct, the upshot is a rather paradoxical situation. Because the deeper we reach for the real, the more aggravated our sense of separation from it then also becomes.

That in this way the intellect is, if we want, engaged in a heroic, potentially tragic, and seemingly unachievable endeavor aimed at finding the naked, fundamental facts of reality within a matrix of perfectly developed sophistications may be nothing new. But living in a so-called information society clearly accentuates the feel of this predicament. To be proper citizens of the information society we must continuously earn our footing. We must maintain and expand an immersive standing within a web that to a large extent appears to us as derivative and abstract, and yet this is also a web that we increasingly rely on as our cognitive backbone. Despite its fundamentally paralocal, astructural, and transpersonal character the machine-powered web of intelligence is becoming the ultimate locus for the most significant and the most intimate facts and details of our lives. For citizens of the information society intelligence is everywhere, and we ourselves, as discrete individuals in time and space, are becoming increasingly hard to pin down.

To state the obvious, even a so-called normal life in a society of this type is a highly complex and demanding project. On the emotional level, the sense of tremendous expansion and empowerment that AI affords us tends to go hand in hand with an equally intense feeling of being overwhelmed and overpowered by factors that are obscure and beyond our reach. It seems safe to assume that, barring global catastrophe, the drive toward an ever more comprehensive application and manifestation of manufactured intelligent agents cannot be halted. The potential payoffs are simply too stunning to be ignored. And indeed, every informed opinion seems to assume[2] that we are just at the doorstep of the AI society, and that the coming decades will bring comprehensive changes—if not transformations that border on, or transgress, the limits of what is humanly conceivable. In other words, we should expect that

[2] For a survey of the expectations of AI-scientific experts, see Ford 2018.

the factors and phenomena that we so far have considered here are only going to increase and intensify in the years to come. This then also means that the paradoxical and rather disconcerting sense of getting both in and out of touch with nature—that uneasy sense of simultaneous expansion and dislocation, depth and estrangement, reality and fantasy—is only going to grow. That would seem to follow as we increasingly invest and entrust our intelligence, creativity, status in society, social life, career, family, passions, dreams, and basic sense of identity in and to the web of intelligence.

§ 2 We Can Never Get What We Want

One way of responding to this problematic ambiguity in our attitude to AI is to seek to personify the intelligent agents that surround us. By making AI look and behave "just like us" to the furthest possible extent we might be able to remedy the lurking queasiness that otherwise accompanies the thought that our ever-present companions are ultimately "just machines." In the service sector at large, efforts are therefore made to produce intelligent machines that can look after us and serve us in ways that display crucial marks of humanity, and hence also emotions. Whether human or machine, we presumably want our caregiver to be well-informed and efficient but also someone who is genuinely concerned about our well-being, feels respect for us, can appreciate a joke, etc. The drive toward affective computing is partly an effort to cater to the latter type of wishes. Nevertheless, although to a certain extent we would like the one looking after us to be "just like us"—namely, to the extent that we can feel "close" and able to share experiences and ideas in a relaxed atmosphere without feeling intimated—we at the same time also wish the caregiver to be in possession of a veritably infinite expertise. Obviously, those two concerns are not easily reconciled. In fact,

they pull apart. On the one hand, I do not wish my nurse or doctor to be simply sweet and harmless. But on the other, how can I trust and confide in a companion that every instant computes and correlates mindboggling myriads of the seemingly disparate pieces of information that the various aspects of my being continuously transmit? Such a caregiver may be a superhuman authority on many of my medical needs, but for trust to be present I need to know who stands before me. And in the case of an AI of this caliber, I clearly do not. For my part, I may then stand in awe and mystery—but most likely also steeped in suspicion, deeply alienated.

This does not bode particularly well for the development of humanoid robots in the service of humanity. What we want from each other as humans, it seems, is nothing that can be formulated as a rational and mutually reinforcing set of qualities. We want things from each other that we couldn't possibly get in a bundle. For example, and as we just saw, many if not most of us want the person close to us to display a frail sense of humanity, a comforting type of fallibleness, so to speak, that makes our own imperfection seem ok. Yet we also want to feel safe and guarded against calamity and mishaps. But who can be sweet and brilliant, cute but infallible, in the same breath?

In the messy world of human interactions we often get by nevertheless, of course. We form friendships and fall in love, resolve conflicts, develop communities, and succeed in taking care of each other despite our mixed bag of ultimately irreconcilable values and desires. But how are we to formulate the objectives and ideals of affective computers if we cannot rely on our common intuitions and ideas? If in the context of powerful AI—and in particular given the possibility of artificial general intelligence (AGI), equaling or surpassing human intelligent capacity—we go by the same barely coexisting and ultimately conflicting set of desires and wishes, the stark tensions between them will likely only be am-

plified. This may lead to highly unappealing and potentially catastrophic consequences. For example, common-sensical drives to rid society of violent crime or petty theft may usher in corrective meassures that drain our lives of privacy or install regimens that may be felt as deeply oppressive. We want safety but we also want to be free—even if these two ideals do not sit well together and must seemingly always be negotiated according to context. With the help of powerfully interconnected AIs, however, our yearning for safety might render free reins to programs of control that once installed will be hard, if not impossible, to reverse. As with the storage of nuclear waste, the storage of collected data can become a matter of life and death for whole populations. Therefore, we cannot just aim to give the people (ourselves included) what they want, because what we want does not hang together as a coherent, sustainable set of objectives.

Having understood this much, we might seek refuge in ethics, and so go searching for a reliable system of higher moral value. Yet precisely because our human aims, values, and desires provide such an idiosyncratic and volatile mix, no matter which set of systematic guidelines we might go for, it seems doubtful whether we, as humans, would be able to tolerate close habituation with machines that operate by strict moral principles. We might here recall the telling fact that most people express a consequentialist orientation when asked whether driverless cars should sacrifice their passengers if that would mean saving the lives of many others on the street. Yet the same people also confess that they would not themselves want to own a car that drives according to such guidelines (Dzikies). Indeed, if potentially super-human moral intelligence operates by guiding lights that shine differently from those of humanity in general the consequences could, from a human point of view, very well turn out to be catastrophic. As Stuart Russel and others centrally involved in the development of AI have argued, if a super-intelligent AI were to take up and pursue any ultimate objective of its

own—no matter how innocent or insignificant that objective might otherwise seem—the ensuing value misalignment between AI and humans would quickly spell global disaster for humanity. As its single ultimate objective comes to override all others, the AI would, so the worry goes, stop short of nothing to see that one aim accomplished, perhaps eliminating all of humanity in the process. Grand ethical theories often involve distinctive versions of the "ultimate good" and hence, if the concerns of Russel and others are real, would seem extremely risky if adopted as the guiding principle for an AI. Myriad unintended consequences become possible such that if, for example, a machine intelligence would be set up to reduce human suffering it might decide to anesthetize and then euthanize an entire population, all the while aiming to protect them from suffeirng.

Humans like us being what we are, we may then conclude, the promise of faithful and endearing, intelligent robotic companions can hardly be fulfilled in any lasting, stable manner. That doesn't mean, of course, that highly capable and emotionally attractive robots cannot be developed. They appear to already exist, as evinced by the rise of social robots in, for example, the health and education sectors. But as long as our idiosyncratic wishes and hopes keep pulling apart, human appreciation for our para- or superhuman companions will at best be fickle, frequently verging on disappointment. Again, rather than the limitations of technology, this continuous shadow of dissatisfaction seems to follow primarily from the fact that what we want and desire from the robots, and from each other in general, is in the end not coherent, meaningful, or practically achievable. A case in point would be our equal requirement of both safety and freedom to pursue individual happiness. Nevertheless, the drive toward ever more endearing and capable robots will continue, pushed on by our disparate wants and wishes.

§ 3 Human Entitlement: Are We Right, Are We Safe?

A disturbing aspect of this development is tied to the fact that we, as mentioned, tend to consider AIs as—well—artificial, and hence not real and actual beings. The very feature of AI that makes us feel dissatisfied with robot companionship—the feature by virtue of which we deem robots subhuman, or subsentient—also gives us a strong sense of entitlement. The otherwise nagging feeling that "there really is no one home" in the manufactured body and mind encourages the conclusion that the robot therefore must exist solely to serve us. Moral considerations on their behalf are therefore simply not relevant, we think. As with some recently developed social robots, the creature before us may seem an awfully polite, sensitive, and well-informed little fellow, but what wrong, we may rhetorically ask ourselves, could possibly be done to a machine? How could one mistreat a toaster or a vacuum cleaner? Any concern, we may then conclude, for the well-being of a robot is simply the result of a category mistake.

Yet on second thought such a conclusion seems less natural and unproblematic. Through history and across the globe, similar conclusions have, as we know, been drawn with precisely the same sense of obviousness and natural justification—and yet today that same thinking may fill us with horror. Simply recalling how attitudes towards certain social groups, ethnicities, or animals, have changed dramatically can give us reason to pause before thinking ourselves naturally entitled, beyond responsibility or blame. What if, for example, the robots of our time, or the near future, should turn out to be somewhat like fish—creatures that in the so-called Western world have largely been thought of as insensitive to pain, but of which the cognitive status is currently under scientific review? What would the ethical consequences be? The closer AI may come to displaying not just intelligence but "artificial consciousness"—especially if at a level indistinguishable from human con-

sciousness—such qualms would obviously become amplified. There might even conceivably come a point at which we would tend to estimate the moral value of super-intelligent AI as superior to our own.

Now, if the incoming scientific evidence should in fact succeed in creating a general consensus that fish possess genuine sentience it may seem morally incumbent on us that we revise a number of commonly held beliefs and practices regarding fish. But even if we don't, and even if fish do in fact feel pain, sticking to the status quo is most likely not going to have any immediate, disastrous consequences for anyone but fish. In the case of a super-intelligent agent in possession of some form of sentience, denying or refusing to acknowledge that sentience would, on the other hand, presumably invite stern corrective measures from the side of the AI.

Moreover, even if there really is "no one home over there"—that is, even if we haven't yet, or couldn't possibly ever, produce a truly sentient being (whatever that might mean)—the consequence of completely denying any sentience in, or moral value of, the artificial Other could still be disastrous. As affective computing keeps taking new strides in the recognition, interpretation, and simulation of human emotions, the response we might receive by treating a virtually emotional super-intelligence as wholly inanimate and morally insignificant might be very similar to what we would get from treating humans that way—even if the machine is in fact "just a machine." Indeed, as their sensors become increasingly sophisticated, such intelligent agents might conceivably respond violently to even our mere *thoughts* about them. In short, regardless of whether AIs can in fact possess sentience of the type we ascribe to ourselves, the safety issues that they raise from the perspective of humans seem very similar.

§ 4 Buddhist Opinions, Buddhist Advice?

The scenario we have depicted here has had three basic themes:

> *Observation 1*: Living with increasingly powerful, interconnected AIs engenders a challenging polarity. We rely on the web of intelligence to obtain solid and factual knowledge, but our very engagements with that web, and even the obtained knowledge itself, typically carries an alienating sense of fluidity, abstraction, and unnaturalness.
>
> *Observation 2:* In the quest for powerful AI we demand irreconcilable qualities and abilities. While we want our robotic companions to be humanoid to the extent that we can feel at home with them we also wish them to be in possession of virtually limitless knowledge and abilities. Thus, we deny robots the limitations and fallibilities that are crucially human characteristics.
>
> *Observation 3:* Our tendency to assume natural entitlement and complete lordship over robots raises important ethical questions and it also spurs potentially lethal behavior. That is the case given complex, higher order subjectivity in machines, but also in the case of barely rudimentary machine sentience, even if it should turn out that genuine sentience in machines is forever precluded by either the forces of nature or the will of the divine (for those who might suppose that divinity restricts sentience to biological entities).

When considering the possibility of a particularly Buddhist perspective on these matters, it seems that there must be plentiful and significant resources to explore. As everyone who has listened to an introductory talk on Buddhist philosophy will know, analysis of the psycho-

physical factors that we associate with self and personhood reveals, from the Buddhist point of view, the absence of any individuality or enduring identity. There are many ways to go, and many paths have been taken, yet Buddhist philosophy and practice nevertheless arguably always proceeds from this basic acknowledgement, or insight.[3]

And indeed, the streaming plethora of diverse views and practices that we now lump together under the rubric of "Buddhism" has been a feature of our world for long. Each distinct tradition has then developed a sustained response to the no-self conclusion and suggested an informed course of action. Moreover, and importantly, all Buddhist approaches aim to integrate their specific variant of the view of no-self with a process for cultivating the ensuing understanding (i.e., what we tend to call "meditation") as well as a supportive mode of conduct. In other words, Buddhism by default seeks to embed some version of a "no-self, no sentient being" perspective within a multi-dimensional and practice-oriented context.

Moreover, the denial of real and enduring persons that follows from the Buddhist analysis of the impermanent factors of existence has—in the face of the intuition that such individuals do indeed exist—given rise to a wealth of ontologies that seek to account for that which "seems-to-be-there-but-is-not-really," all the while seeking to stay clear of both reification and denial of what is presented and constructed in experience. In this way, Buddhist traditions have for millennia operated under the constraints of an analytic denial of real sentient beings, while nonethelss taking this as the framework for developing deep ethics and

[3] Even the ancient Pudgalavāda appear to have formulated their doctrine of the "inexpressible" (Skt. *avācya*) self that neither lasts and nor ceases, etc. as a way to provide a foundation for karmic ripening where there is no eternal soul or lasting self. Similarly, when contemporary Buddhist scholars insist on a conventional, empirical, or psychological self they do so against the backdrop of the classic deconstruction.

rich currents of epistemology. In the Mahāyāna in particular, knowledge of no-self is associated with a fearless capacity of universal care for the infinity of apparent, suffering beings.

As we seek to understand and respond to the challenges raised by AI and VR, it seems that we should be looking for responses that are not only open to, but in fact *informed* by, an understanding of the world and its beings as merely apparent, virtual but not real, yet whose experience and potential suffering nonetheless matters. We need perspectives that can make sense—preferably common sense—of the artificial and the virtual, and at the same time suggest ways of being and acting that are ethically wholesome and aesthetically pleasing. If our perspective comes out lacking, or lopsided, in any of those regards the ensuing imbalance risks setting off a downward spiral. The dangers of mismanaged or misunderstood AI are, of course, serious and real. Now, the various Buddhisms of the world are typically used to operating simultaneously within the spheres of ontology, ethics, and aesthetics. As far back as we can see, this is how Buddhism has been getting by: An ontology of the virtual and artificial is integrated within a doctrine of transpersonal care, and presented as an attractive, satisfying approach to whatever challenges may be current. If there ever was any deeper value to the Buddhist perspectives, I cannot help but think that valuable resources for the information society should be aplenty there.

From what we may call "a Buddhist perspective," there can hardly be any difference in the status of humans, animals, and robots if what we are looking for is a singular, permanent, and independent individual. That is because in all three cases such a being is wholly absent. If, alternatively, we ask whether humans, animals, and robots may possess intelligence, it seems hard to deny that they can, and do. Moreover, when machine intelligence is termed artificial, such a qualifier would, according to Buddhist understanding, likely apply to the human intellect as

well, at least if we follow Nāgārjuna's classic critique of the fabricated vs. the natural.[4] The question of a fundamental difference between the axiomatic three classes of beings—humans, animals, and robots—could thus, from this Buddhist perspective, only be in terms of the presence or absence of sentience. And on the characteristics and functions of sentience the diversity and complexity of Buddhist opinion is again remarkable. As with the view of no-self, understanding the nature of sentience—if indeed there is one—is key to Buddhist knowledge, and so resources on this topic abound.

But let us here perhaps just contend with noticing that while Buddhism denies the ultimate reality of enduring or even impermanent sentient beings, this denial is combined with a doctrine of strong, at times passionate, love for the world and its beings. Although this might seem a mystery—since according to the Buddhist view there are no beings of substance—Buddhism, and Mahāyāna in particular, takes the recognition of the nonexistence of enduring, action-controlling sentient beings and the expression of self-less, compassionate love to be two sides of the same insight. The recognition of *anātman*—no-self—serves, we might suggest, to cut through blind wants and desires. Thereby painfully unachievable demands—such as the irrational requirement that our companion should be both cozily flawed and perfectly capable—may be painlessly relinquished. Such relinquishment allows, in turn, for the flourishing of cognitive and affective qualities that transcend the fixtures of schematized individuality—natural intelligence, if we want. And even the idea of such intelligence as a web is actually not foreign, at least not to tantric Buddhism where it is a central concept. But we are clearly getting far ahead of ourselves here, and I shall curb my admittedly rather simpleminded enthusiasm. All that this essay aims to suggest is that

[4] *Mūlamadhyamakakārikā* XV:1-2.

Buddhist understanding and ways of life should provide useful resources for an intelligent and ethically sustainable cohabitation with artificial intelligence. Let me conclude with a stanza from the Maitreya-ascribed *Mahāyānasūtrālaṃkāra* that according to the commentarial literature presents the distinctive qualities of the wisdom of great compassion. Characteristically dense and packed with seminal meanings, the stanza culminates with a haunting claim:

> If it is not equality, eternity, superior intention,
>
> Means for accomplishment, freedom from desire, and absence of focal point,
>
> Then it is not love,
>
> And where there is no love, there is no awakening.[5]

Not only in terms of how we perceive and interact with so-called artificial beings, but also for the ongoing work with developing them, a declaration such as this, and the context from which it comes, may help suggest safe and meaningful avenues.

Works Cited

Dizikes, Peter. "3 Questions: Iyad Rahwan on the 'psychological roadblocks' facing self-driving cars." *MIT News*, http://news.mit.edu/2017/3-questions-iyad-rahwan-psychological-roadblocks-facing-self-driving-cars-0911.

[5]*Mahāyānasūtrālaṃkāra* XVIII.35: *na sā kṛpā yā na samā sadā vā nādhyāśayād vā pratipattito vā / vairāgyato nānupalambhato vā na bodhisatvo hy akṛpas tathā yaḥ //*. The translation is based on the Tibetan (Tōh 4020): *mnyam dang rtag min lhag pa'i bsam pas min // sgrub pa'i sgo nas 'dod chags bral ba dang // mi dmigs pas min gang de brtse min te // de bzhin brtse med gang de byang chub min //*

Ford, Martin. *Architects of Intelligence: The Truth about AI from the People Building It*. Packet Publishing, 2018.

Maitreyanātha. *Mahāyānasūtrālaṃkāra.* n.d. Sanskrit edition in Levi, Sylvain. *Mahāyāna-Sūtrālaṃkāra: Expose de la Doctrine du Grande Vehicule.* Librarie Hononoré Champion, 1907. Tibetan translation in *Theg pa chen po mdo sde rgyan*. Tōh. 4020.

Nāgārjuna. *Mūlamadhyamakakārikā*. n.d. Sanskrit edition in de Jong, Jan Willem (ed.). *Mūlamadhyamakakārikāḥ*. Adyar Library and Research Centre, 1977.

Russel, Stuart. "Provably Beneficial Artificial Intelligence." In *The Next Step: Exponential Life.* BBVA-Open Mind, 2017. 175-92.