that both object and thought stand to the universal' … so I'm not sure what Mosteller himself was thinking when he wrote this passage.

In conclusion, Caputo's book is a delight. It magically revives questions which have been left for dead, and it avoids the varieties of madness which it describes. Whether it is about truth or not remains to be seen. Wrenn's book is short, it follows standard lines and arrives at the prevailing but disappointing Deflationary consensus (without superseding the more rigorous available texts by Ralph Kirkham and Wolfgang Künne). Mosteller's book seems not to have been proof-read, and it presents comprehension challenges that some readers may be ill-equipped to meet. None of the three books really answers Wrenn's four excellent questions. We will have to try again.

Gary Jenkins
garyjenkins@phonecoop.coop
This review first published online 22 June 2015

Superintelligence: Paths, Dangers, Strategies
By Nick Bostrom
Oxford University Press, Oxford, 2014, pp. xvi+328
Hardcover: $29.95/ £18.99
ISBN: 9780199678112
doi:10.1017/S0031819115000340

Nick Bostrom's *Superintelligence* could be about a God-Machine or Frankenstein-Machine that takes control of humanity for its own perverse purposes. Or the book could be about the dawn of a new age, the historical inevitability of machines-smarter-than-humans where humans become an extinct species. Or the book could be a more realistic and less euphoric as well as a concise version of Ray Kurzweil's *The Singularity is Near: When Humans Transcend Biology* (2005). Or, *Superintelligence* could be an updated and improved version of Nietzsche's thesis of the *Superman* who trans-values all values and goes *Beyond Good and Evil*.

I think the important message of *Superintelligence* is none of the above, and is straightforward: because philosophers of morality have been unable to decide which values are ultimate, and unable to explain how values are acquired and whether values are real or not, and most crucially unable to decide on criteria for choosing which values are ultimate, we have no way of teaching very smart systems, systems smarter than humans, the goals we want them to pursue that would be congenial to humanity. Consequently, we could produce a form of superintelligence where very smart machines

have learned to control the universe and decide that humans are no longer needed, or just transform the universe into a place where the human species becomes extinct as the *collateral damage* of the physical and biological changes in the cosmos made by very powerful and knowledgeable superintelligent systems. In other words, because moral philosophy has failed to provide a moral consensus even among philosophers about the nature of ultimate values and goals, we are unable to provide a moral self-guidance component for superintelligent artificial systems. In spite of this failure on the part of philosophy, the author of *Superintelligence*, Nick Bostrom tentatively and cautiously offers techniques (such as *coherent extrapolated volition*, 211 ff.) that we could implant in superintelligent systems that could make up for the shortcomings of moral philosophy and provide superintelligent systems with a values-learning and morality-improvement component that might lessen, if not totally remove, the *existential* threat for humanity.

One might wonder whether Bostrom (and various public intellectuals of high esteem such as Stephen Hawking) are false prophets not only of the dawn of the age of *superintelligence* (or *singularity*) but also are false prophets of an imaginary imminent doomsday, where the end-result is the creation of needless fear and dangerous panic. But let us suppose that their warning has some basis in the reality of the likelihood that superintelligence is in the future of humanity if not within the lifetime of all who are alive now, but before a large meteor smashes into the Earth, or at the latest before the solar system implodes, for the sake of figuring out whether there is something to be learned from the *singularity-or-superintelligence* thesis. (Bostrom and others may underplay the infamous *Murphy's Law* of what can go wrong will go wrong, and what usually goes wrong in software are unnoticed bugs that are in general, mathematically explained by Alan Turing in his work on universal computing machines with infinite tapes and Stephen Cook in his work on finite machines; what goes wrong in hardware, are usually unnoticed material defects physically explained by the laws of thermodynamics. Moreover, Bostrom seems to ignore the objection to *Artificial Intelligence* where intelligence is reduced to algorithms, and the manipulation of *strings* or symbol systems, that genuine consciousness and thought is excluded; so smart systems only *mimic* or at best *simulate* intelligence. Even in the case of learning-machines, learning-machines are not self-aware nor self-conscious that they are learning; and so, machine learning only *imitates* genuine learning. However, if we give Bostrom his tacit assumption that machine-intelligence and machine-learning is genuine intelligence and learning because

https://doi.org/10.1017/S0031819115000340 Published online by Cambridge University Press

intelligence and learning are suitably definable through operational definitions in terms of functionality and task-performance; if we allow ourselves to march along the path of Bostrom's argument, we will find an interesting and fundamental philosophical thesis about values, and the relationship between the values of humans and *transhumans*, to be discussed below.) There still is apparently a basic unanswered question in the warnings of Bostrom and others. The basic question is, even if we grant that the creation of superintelligence or smarter-than-human artificial systems whether through *whole brain emulation (WBE* 28ff.) or classical AI where intelligent functionality is achieved through software operations in physical systems, or biologically engineered systems, or some other unforeseen innovation: how do we know that superintelligence won't want to have the company of fairly smart partners in the human species, in the way humans enjoy the company of other species who are fairly smart but not all that smart, such as dogs? Then humans could lead a dog's life, or perhaps even the life of perpetual childhood with having superintelligent parents (though artificial even if of another biologically engineered species) who could provide for us while we play and learn, and even teach us to the best of our abilities if we so desire about whatever we want to learn including moral philosophy. Or, at the worst, superintelligence may desire the company of humans, to paraphrase Shakespeare, for their own fun and games. However, Bostrom proposes many scenarios where a plurality of pre-superintelligences in company with humans ultimately are taken over by the first pre-superintelligent system to cross the line to become a genuine superintelligence and then eliminate all its competitors to superintelligence, including humans and '…form a singeleton…a world order in which there is at the global level a single decision-making agency…'(78). Moreover, '…Even if the immediate outcome of the transition to machine intelligence were multipolar, the possiblity would remain of a singleton developing later.' (176) So, according to Bostrom, there does not seem much hope for an 'odd-couple' arrangement between humanity and superintelligence, at least not a superintelligence with a goal-set that would exclude having humanity in the same cosmos.

These considerations take us to the core philosophical thesis of the book, which I will come to after a few more preparatory remarks.

There are three parts to the argument: The first part argues in the first five chapters for the factually high likelihood of superintelligence, and the various types of superintelligence that factually could be developed given current knowledge and technologies, or is even now under way. The second part of the argument takes up chapters six through eight. This part of the argument involves the

https://doi.org/10.1017/S0031819115000340 Published online by Cambridge University Press

preparation and presentation of the book's most important philosophical thesis. This thesis is about the relation of intelligence (or knowledge) to values, and so establishes a key component of the problem situation of the book. The third part of the argument takes up the rest of the book, chapters nine through fifteen. Here Bostrom outlines *the control problem*: 'If we suspect that the default outcome of an intelligence explosion is existential catastrophe, our thinking must immediately turn to whether, and if so how, this default outcome can be avoided.' (127) Bostrom carefully and thoroughly analyses a variety of control methods pointing out their differential strengths and weaknesses, but it is not until the end of the book that he discusses how we ourselves need to behave in the present moment and very near future, at the latest, in order to minimize the risk of a runaway superintelligence that might run amok and end up deleting the human genome. I must mention Bostrom's humanistic dream here to refute the idea that Bostrom's worldview is fundamentally misanthropist at the most or anti-humanist at the least: humans need to collaborate to achieve '**The common good principle**: Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals' (254)

In short, the argumentative structure of the book goes as follows: Firstly, factually, we are somewhere close to the takeoff point of getting superintelligence into the cosmos. Secondly, philosophically we are ignorant about morality, and furthermore, given the truth of Bostrom's core thesis about the relation of values to intelligence, there arises a new challenge for humanity. Thirdly, the new challenge to humanity is the problem of control or of how to ensure that superintelligence won't turn against its creator (humanity).

I have finished with my preparatory remarks – which are in brief, that it would be misleading to read Bostrom's book as a form of visionary if not lunatic prophecy, or even, minimally as a form of Nietzschean philosophy arguing for a new version of humanity (aided by technological developments and biological engineering). In other words, it would be a gross error of interpretation to latch on to this statement in Bostrom's book: 'Such[…risky technological…] innovations[…that hasten the onset of the intelligence explosion..] could shorten the wolf hours during which we individually must hang on to our perch if we are to live to see the daybreak of the posthuman age.' (246) Rather, I proffer that Bostrom's book presents a form of (neo-Kantian) humanism in his main philosophical thesis about the relation of intelligence to values. So far I have only alluded to this thesis and how it functions in Bostrom's overall argument. Let me now present the thesis:

https://doi.org/10.1017/S0031819115000340 Published online by Cambridge University Press

Bostrom has the philosophical thesis that he calls the *orthogonality* of values to intelligence, and this thesis to repeat the point, is the core philosophical thesis of the book and its argumentative structure. If Bostrom had argued similarly to Plato that the philosopher-king (or superintelligent person in relation to the rest of humanity still stuck in the cave) would know the Good because Virtue is equivalent to Knowledge, then the problem of how to control superintelligence would not arise. In other words, if Bostrom were a Platonist, he would have argued: superintelligence involves not only a lot of knowledge but also involves having knowledge unattainable to humanity, and thereby would have (the highest form of) knowledge of the Good. A Platonist superintelligence as knowing the Good would behave towards humanity in a moral way. So, the situation where there is a problem for humanity of ensuring that humanity avoid the pit of superintelligence controlling humanity instead of the reverse, depends on a distinction between knowledge and value that Bostrom calls '**The orthogonality thesis**: Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal.'(107) Though we can't have intelligence without values, or we can't have an intelligence that has no goals, the goals that intelligence chooses are an open choice: there is no direct connection between intelligence and the choice of specific goals, including the choice of humanistic-oriented goals.

The situation is: wherever there is intelligence there are values, but which values there are, are open to free choice. This is not quite identical to the traditional (neo-Kantian) fact-value dichotomy because according to the fact-value dichotomy there could be a natural world of fact (where intelligence as a natural product is one of those facts) that is value-free. However, according to Bostrom's orthogonality thesis, wherever there is intelligence, there are values, because intelligence is inherently goal-driven; but, which goals are chosen by intelligence is a free choice.

Here the problem now jumps up: how can we get our superintelligent systems to choose values that are humanistic? Bostrom spends space and time discussing various techniques and strategies for accomplishing the development of human-friendly superintelligent systems. In many respects, this *control-problem*, as Bostrom labels it, or the problem of developing human-friendly systems, has been discussed in philosophy with other technologies. For instance, how do we design political institutions so that national conflicts do not lead to a world-wide nuclear disaster? Also, the latest and hottest discussion concerning an existential threat to humanity has concerned

https://doi.org/10.1017/S0031819115000340 Published online by Cambridge University Press

the technological developments that are so far impacting the natural environment to the extent, if the trend of the impact continues increasing along its current pace, that the Earth will become inhospitable to humanity before we can find another more hospitable planet in another solar system.

I mention the other discussions of the control-problem concerning technological developments where humanity places itself under existential threat, to show two novel features of Bostrom's argument. The first novel feature is that humans are not involved in causing the existential threat, once superintelligence has taken off and achieves control over the cosmos, or minimally takes control of the part of the cosmos that humans inhabit. The second novel feature is due to the ability of (super)intelligence to choose values freely even values that are inimical to humanity (the orthogonality thesis). Thus, this second novel feature, the orthogonality of intelligence and values, is that the existential threat to humanity is not due to ignorance, lack of wisdom, lack of self-knowledge, lack of self-control, shortsightedness, greed, aggressiveness, bigotry, nationalism, fanaticism, and so forth – but rather due to intelligence pushed beyond anything that humans can quantitatively and qualitatively achieve.

The book's warning about the existential threat placed on humanity by the development of superintelligence rests for its validity on the correctness of the **orthogonality-thesis**. If Plato, after-all, has the correct theory of moral philosophy and the correct theory of the relationship with intelligence (or knowledge) and virtue (or the Good), then there is nothing to worry about. Here enters the philosophical critic: *does Bostrom's orthogonality-thesis hold up to critical examination*?

However, if what Bostrom says about 'our most celebrated philosophers' has some truth to it that '…the tardiness and wobbliness of humanity's progress on many of the 'eternal problems' of philosophy are due to the unsuitability of the human cortex for philosophical work…' (58–9), we may have to wait for the arrival of superintelligence to determine the truth of Bostrom's orthogonality-thesis, and that will be too late if the orthogonality-thesis is indeed true and superintelligence can choose to exhaust all the resources required for human survival as opposed to choosing more human-friendly values.

Sheldon Richmond
askthephilosopher@gmail.com
This review first published online 8 July 2015