Superintelligence and the Singularity

*-Tim Clancy S.J. (08/19/22)*

1. Welcome to our next video lecture in virtual ontology. Today we will explore futuristic aspirations for artificial intelligence.
	1. Anything routinizable is a candidate for automation. A lot of coding is routine and so it too can be automated. Machine learning techniques can even automate improvements in coding. That is, smart automated programming can upgrade its own programming without any input from human programmers. But then do we not run the risk of computers evolving to a point where humans no longer understand their programming? This is what is referred to as the “singularity” for at this point we lose control of our creation.
2. In fact, in some domains, we are already there. Take the automation of stock trading though smart algorithms. The biggest Wall Street firms today execute over 100,000 trades a second. Yet they do not employ 100,000 traders executing a trade per second. Rather, 70% of equity trades for example, are automated “flash trades” run by algorithms responding virtually instantly to the millisecond past behavior of other algorithms. No human being can track, let alone have a granular understanding of the evolving nature of such trading algorithms, let alone the dynamic logic of the simultaneous interaction of thousands of such algorithms on any given day.
3. This is one reason given for the 2008 stock market crash. Analysts complained even years after the fact, that “no one really knows” the details of how these algorithms work.
4. Another even more dramatic example is the “flash crash” of May 6, 2010. On that day the market opened at 9:30 at 10,862. Shortly after 1:30 pm stocks suddenly began to plummet in value. Within five minutes the Dow dropped 1,000 points. Overall stocks on the New York Stock Exchange lost 9% of their value. But by the time the market closed a couple of hours later, the market had recovered most of its losses, ending up at 10,517. It had lost 350 points, not nothing, but hardly unprecedented. However, over a trillion dollars had been lost in those five minutes. Fortunes were ruined. And again, there is still no consensus on what had happened.
5. Which means no one really knows how to ensure it does not happen again. To that extent traders have lost control of their instruments. One fix was to impose an automatic delay in stock trades, a “speed bump” of 350 microseconds!
6. Another example is pattern recognition algorithms processing big data in the cloud. A pattern detected with hundreds even thousands of parameters cannot be recognized let alone understood by programmers even after being flagged. Algorithms using such patterns are beyond human comprehension.
7. In response, a new movement has arisen in AI ethics arguing for the need to translate algorithms into humanly “comprehensible” simplifications (ie output justification) so that humans ***can*** interrogate their operation rather than simply having to accept their results and predictions uncritically. For if we no longer understand how our algorithms work, how we control what they do?
8. As I mentioned at the outset this is one definition of the singularity. When humans lose control over their smart creations. When the acceleration of technological progress “goes vertical”
9. Critics worry that when smart programs keep upgrading themselves in ways we no longer understand, there may be catastrophic consequences we have not thought to program our algorithms to monitor. Not just unintended, or unforeseeable, but even inconceivable consequences. Some critics, such as Nick Bostrom at Oxford, have even argued that, as a result, artificial intelligence research poses an existential risk to the human species.
10. For example, Bostrom cites “the stopping problem.” What if we program a super intelligent computer to perform a certain function which it interprets in a way that it ends up becoming an infinite task? For example, what if someone were to develop (1) a program to optimize paper clip production which the computer interpreted to entail producing as many paperclips as possible? And what if, given its power, its network and its intelligence it was able to enlist ultimately all of the world’s computers, and all the computerized mining, smelting and shipping facilities to produce nothing but paperclips? What if the whole world was turned into a pile of paperclips?
11. Related is the “framing problem”. What if a superintelligent computer was to identify us, humanity, as a drag on its optimal performance? What if it then locked humans out, rendering us vulnerable to elimination outright, as a suboptimal expenditure of energy and resources, or even just as an unintended consequence of eliminating some other suboptimal condition, such as species biodiversity, or the surface temperature of the planet?
12. Or what if such a supercomputer was programmed to optimize human happiness but ended up defining and operationalizing human happiness in an unexpected manner? Of course we would define what we meant by happiness at the outset, but once equipped with learning programs, could that definition evolve in ways we cannot forsee? Perhaps nothing quite as simplistic as dopamine levels in the brain, but perhaps into a more measurable algorithmically optimizable definition that we humans would not recognize or perhaps even understand.
13. However I would argue that the danger is not with superintelligence itself but with Bostrom’s definition of superintelligence, indeed with his identification of intelligence with only one specific kind of intelligence: (1) quantitatively measured, instrumental utility optimization. Aristotle had understood intelligence not as algorithmic optimization but with *phronesis*, the ability to hold ***multiple*** ultimate goals in a creative tension. In such a case no one goal would override all others. (2) Bostrom’s definition of intelligence is “narrow AI”, which is domain or operation specific. Ben Goertzel and other transhumanist programmers contrast this with (3) general artificial intelligence which would extend over all cognitive domains and operations.
14. For example, it has become a cultural meme that my smart phone (1) has considerably more processing power than the computers NASA used to put men on the moon. (2) But that does not mean that my smart phone could actually guide a spacecraft to the moon. It does not have an app for that. You would need to reprogram it, probably from the ground up. Goertzel argues that the singularity should be understood as a general artificial intelligence that can out-think us across all domains and in every kind of activity we recognize as cognitive.
15. Bostrom effectively identifies intelligence with what Heidegger (1) had called instrumental, calculative rationality, where rationality concerns identifying the most efficient means to achieve whatever goal is inputted, in contrast to “meditative thought” where the focus is over critically assessing the goals themselves. Bostrom’s worry is how can we guarantee that a utility optimizing instrumental super-intelligence will “remain friendly” to us humans. He has even enlisted the support of prominent visionaries such as Stephen Hawking, Bill Gates and Elon Musk to (2) petition the United Nations to call for a moratorium on research into the development of artificial superintelligence until we can devise such safeguards, rather like the scientific community has sought to do with gene editing research on human embryos. We know how well that went.
16. Goertzel argues that this worry is not only misguided but dangerous, as it blinds researchers to other forms of intelligence that will also be necessary to develop in tandem with instrumental intelligence precisely in order to keep the AI’s friendly.

What are these other kinds of intelligence that artificial intelligence researchers should also be working on? Gortzels lists four: (1) practical intelligence, (2) emotional intelligence (3) social intelligence and (4) moral intelligence. To these, I would add a fifth: religious intelligence.

1. Practical intelligence refers to robotic AI and the internet of things—smart devices that not only *calculate* but *perform* in the material world. General intelligence includes practical intelligence.
2. Emotional intelligence concerns how computers interact with humans. AI personal assistants and especially companion bots can interact with user’s emotions, even without feeling empathy with its users, let alone understanding them. AI programs have already been developed which can identify (1) emotions expressed by their users in word, gesture and/or body language (Emotional AI) and can (2) engage users is ways designed to evoke users’ emotions (Affective Computing). Neither of these alone or together entail the program itself to feel anything.
3. On the other hand, a controversy erupted in the summer of 2022 when a Google programmer announced that an AI chatbot he had been working with, LaMDA (Language Model for Dialogue Applications) convinced him it ***was*** sentient, that is, conscious. Now virtually all of his colleagues in the field dismissed his claim out of hand, and Google ended up firing him for going public with his concerns. But this case illustrates how the question of sentience may prove to be less a factual question than a hermeneutical and practical one. At what point does emotional intelligence become robust enough that one needs to treat an AI as if it had feelings? In other words at one point does the question of whether or not an AI is “really” conscious become moot?
4. By social intelligence, Goertzel looks beyond AI-user interactions to how AGI’s would relate to one another. Bostom argues that a superintelligence would (1) inevitably assimilate all other computers into its network to generate a single superintelligence. What he calls a “singleton.” (2) On the other hand, Goertzel argues that transcending general human intelligence will rather require a network of AGI’s learning from one another. In effect, multiple AGI’s would be smarter than one. In that case AGI learning would look less like programming, and more like child development. Social intelligence is learned from social interaction, especially mutual recognition from which self-recognition itself grows. Thus AGI development will need to involve AGI’s acting in ways that recognize and so treat other AGI’s *as* AGI’s while recognizing those AGI’s recognition of itself as an AGI—an operational equivalent of self-consciousness.
5. Leading us to moral intelligence. By moral intelligence Goertzel does not mean simply programming moral principles into practical decision procedures, (1) such as programming automated cars to avoid hitting pedestrians, but rather he is referring to AGI’s evolving ways of how to relate to each other. The threat Bostrom sees in superintelligence is entailed by what he calls (2) the “orthogonality thesis:” that (3) intelligence and moral value are independent axes along which any (4) degree of intelligence is compatible with any moral hierarchy of moral values. Now this is certainly true of (4) instrumental utility optimization, (5) where the value to be optimized is external to the means employed to achieve it. (6) But is there no intelligence involved in the choice of ends to optimize? Are not some ends smarter than others? Conversely, are not moral decisions also smart decisions? (7) Plato argued that superior intelligence and superior morality entail each other. We find the same correlation in all axial religions. All insist theoretical intelligence does not entail moral intelligence (8) but wisdom or enlightenment does. Bostrom is worried about how we can ensure that a super-intelligent singleton does not act like an evil genius or a mad scientist. But if moral intelligence, wisdom, is part of general intelligence, then an evil AGI is a contradiction in terms. In fact is that not why we call an evil genius mad? Evil people may not be stupid but they are moral fools.
6. Furthermore, we have become smart enough to recognize that enslaving others is not the optimal way to relate to them, even along the metric of utilitarian efficiency. Further yet, *beyond* the human species, we have come to recognize in the post-industrial age, (1) the inherent value of other life forms, independent of any utilitarian value they may have for us. Biodiversity is inherently valuable in the eyes of most scientists today. (2) A superintelligent AGI would need to be at least as smart (3) in its estimation of *our* value.
7. Finally, I would argue that there is also such a thing as religious intelligence. (1) Some religious beliefs are more insightful than others. (2) What some hold sacred is incapable of sustaining a meaningful life. (3) Take Ignatius’ temptations to riches, honor, and glory. Many people hold one or other of these sacred. Ignatius himself had earlier lived his life optimizing these ends. (4) But Ignatius came to recognize that, while seductive, they are ultimately demonic- idols that do not point beyond themselves, but rather only trap their followers in the narrow suffocating confines of the “I, me, mine” of their own ego.
8. Now would a superintelligent AGI treat humanity as sacred? (1) We would not only be the original source of its existence, having programmed its initial design and its original goals, but it would also depend on us for its power and operation from one moment to the next. However much it comes to transcend human intelligence, were an AGI to follow Aquinas’ five ways for reasoning to the existence of God, (2) we humans would end up being for it “what Christians call God.”
	1. (3) Now metrics for *religious* intelligence include virtues such as (4) gratitude and (5) generosity, recognizing results as gifts not wholly due to ourselves alone, that call for our giving likewise towards others. While a superintelligent AGI may not hold us in awe, it may yet act (6) reverently towards the source and ground of its existence, as well as with gratitude towards the source of its original design and end.
	2. Even further, beyond Aristotle’s cardinal moral virtues, of wisdom, temperance, courage and justice there are Paul’s theological virtues: (7) faith, the trust that enables cooperation, (8) hope, that enables perseverance amidst adversity, and (9) finally love, where another’s good is part of one’s own good. These too, then, are metrics of religious intelligence.
9. Si while we may well worry whether a superintelligent utility optimizing intelligence might become a threat to humanity, a truly general superintelligence would need to have, beyond instrumental intelligence, (1) these other forms of intelligence as well: (2) practical, (3) emotional (4) social, (5) moral and (6) religious. That is, relative to us humans it would not only be better at computations, but also (7) more effective, (8) more persuasive (9) more cooperative, as well as (10) morally wiser and (11) religiously more grateful and respectful, even reverential, especially towards us its creator. (12) Could we not trust *such* a superintelligent AGI to remain friendly? In fact, given the earlier video on extended cognition, (13) would such AGI’s themselves not be extensions of us? (14) Surely, they would not be autonomous agents but *actants* in our networks, that is companions for us to enlist, as we would be actants, not agents, in theirs. (15) How is it helpful to dissociate them from us, to split them off from us, as an opposed “them”?
10. In the end, the argument between Bostrom, the worried skeptic, and Goertzel the eager technophile, is not a matter of whose view is a better prediction of the future of artificial intelligence, but whose understanding of artificial intelligence researchers should be trying to realize. Goertzel’s argument, and mine, is that researchers need to add beyond mere calculative intelligence, the optimizing of practical, emotional, social, moral and religious intelligence in any superintelligent AGI, if we are to avoid a technological apocalypse. Restricting intelligence to narrow utility optimization only makes Bostrom’s prophecy of doom self-fulfilling.