Uziel Awret

Introduction

In his 1965 article 'Speculations Concerning the First Ultraintelligent Machine' statistician I.J. Good predicted the coming of a technological singularity:

> Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

The term 'singularity' was introduced by the science fiction writer Vernor Vinge in a 1983 opinion article. The underlying idea has always captured the imagination of science fiction writers from John Campbell's 1932 short story 'The Last Evolution' to Robert A. Heinlein's 1952[1] essay 'Where To?', Asimov's 1956 'The last question' and many more recent works. There has also been discussion by mathematicians, AI researchers and futurists. Until recently, the subject has not been as popular with philosophers.

This special interactive interdisciplinary issue of JCS on the singularity and the future relationship of humanity and AI is the first of two issues centered on David Chalmers' 2010 JCS article 'The Singularity, a Philosophical Analysis' (available online at www.imprint.co.uk/jcs.html). These issues include more than 20 solicited commentaries to which Chalmers will respond later this year. To quote Chalmers:

> One might think that the singularity would be of great interest to Academic philosophers, cognitive scientists, and artificial intelligence

Correspondence:
Email: awretu@trinitydc.edu

[1] See Damien Broderick's article on the singularity and science fiction in this volume.

> researchers. In practice, this has not been the case. Good was an eminent academic, but his article was largely unappreciated at the time. The subsequent discussion of the singularity has largely taken place in non-academic circles, including Internet forums, popular media and books, and workshops organized by the independent Singularity Institute. Perhaps the highly speculative flavour of the singularity idea has been responsible for academic resistance to it. I think this resistance is a shame, as the singularity idea is clearly an important one. The argument for a singularity is one that we should take seriously. And the questions surrounding the singularity are of enormous practical and philosophical concern.[2]

It is fair to say that Chalmers is the first to provide a detailed comprehensive philosophical analysis of the idea of the singularity that brings into focus not only questions about the nature of intelligence and the prospects for an intelligence explosion but also important philosophical questions about consciousness, identity and the relationship between facts and values.

At the end of the 2010 Tucson consciousness conference during one of the celebrated 'end of consciousness' parties, Chalmers and I discussed his plenary talk on the singularity and agreed that it was well-suited for a special edition of JCS. The idea was to solicit commentaries on his target article from philosophers, AI researchers, science fiction writers, futurists, cognitive scientists, biologists and others to which Chalmers would respond. This project is also commensurate with the JCS credo of exploring controversies in the science and the humanities.

We received many invited and submitted commentaries, enough to fill two special issues. This first issue includes articles by Nick Bostrom and Carl Shulman, Sue Blackmore, Damien Broderick, Barry Dainton, Dan Dennett, Ben Goertzel, Susan Greenfield, Robin Hanson, Francis Heylighen, Marcus Hutter, Drew McDermott, Juergen Schmidhuber, Frank Tipler, and Roman Yampolskiy. The second issue, to be published later this year, will include articles by Igor Aleksander, Richard Brown, Ray Kurzweil, Pamela McCorduck, Chris Nunn, Arkady Plotnitsky, Jesse Prinz, Susan Schneider, Murray Shanahan and Burt Voorhees.

Chalmers paper is divided into three parts, the likelihood of the singularity, negotiating the singularity, and the place of humans in a post singularity world with a special emphasis on uploading.

---

[2] Chalmers mentions some exceptions to this academic neglect including Bostrom (1998; 2003), Hanson (2008), Hofstadter (2005), and Moravec (1988; 1998).

I will use a synopsis of his paper to present short descriptions of the different contributions to this volume.

Chalmers' basic argument for the singularity is:

1. There will be AI (before long, absent defeaters).
2. If there is AI, there will be AI+ (soon after, absent defeaters).
3. If there is AI+, there will be AI++ (soon after, absent defeaters).

---

4. There will be AI++ (before too long, absent defeaters).

AI is human-level intelligence, AI+ is greater than human intelligence and AI++ is much greater than human intelligence (standing to humans as humans stand to ants). 'Before too long' means within centuries while 'soon after' means within decades or years. Defeaters are defined as anything that prevents intelligent systems from realizing their capacities to design intelligent systems.

Chalmers analyses the first three premises separately describing them accordingly as the equivalence premise, the extension premise and the amplification premise.

The equivalence premise (we will construct AI as intelligent as ourselves) includes the brain emulation argument and the evolutionary argument. The emulation argument claims that:

(i) The human brain is a machine.

(ii) We will have the capacity to emulate this machine (before long).

(iii) If we emulate this machine, there will be AI.

(iv) Absent defeaters, there will be AI (before long).

Neuroscientist Susan Greenfield argues against both premise (i) and (ii) and attempts to provide what she calls a reality check arguing that the brain is non-computational and *that whilst the hypothetical scenario of neuron substitution is conceptually logical and plausible, in reality it's meaningless and unhelpful*. Greenfield also feels that consciousness is crucial for values, understanding and 'wisdom'.

AI researcher Francis Heylighen also seems to reject both premises (i) and (ii) embracing the embedded paradigm in which the brain does not simply crunch symbols and is inseparable from its immediate environment.

In the second volume philosopher and cultural theorist Arkady Plotnitsky holds that microphysical processes cannot be simulated arbitrarily closely and that the emulation argument and the evolution argument fail to convince us that we will have AI soon.

The evolutionary argument proceeds as follows:

(i) Evolution produced human-level intelligence mechanically and non-miraculously'.

(ii) If evolution produced human-level intelligence, then we can produce AI (before long).

(iii) Absent defeaters, there will be AI (before long).

How difficult is it for evolutionary mechanisms to produce intelligence similar to ours? Carl Shulman and Nick Bostrom address this question by salvaging the evolutionary argument from the 'observation selection effect' objection. They do so by combining arguments which are based on relevant examples of terrestrial convergent evolution with probabilistic arguments that are based on the 'sleeping beauty paradox' concluding that the evolution of human level intelligence on an earth type planet is not exceedingly improbable.

Economist Robin Hanson agrees with all three premises but claims that concluding that human level AI is near is based less on Good's recursive argument with its ensuing intelligence explosion and more on the extrapolation of general historic and economic trends that are clearly exponential.[3] Hanson also holds that the relevant parameters that should be traced in the context of an intelligence explosion are not those of individual systems, whether biological or artificial, but rather more collective 'cognitive' feats. This leads us to the extension premise leading from AI to AI+.

(i) If there is AI, AI will be produced by an extendible method.

(ii) If AI is produced by an extendible method, we will have the capacity to extend the method (soon after).

(iii) Extending the method that produces an AI will yield an AI+.

(iv) Absent defeaters, if there is AI, there will (soon after) be AI+.

Three extendible methods are put forward: direct programming, machine learning, and artificial evolution. AI researcher Drew

[3] See his article in this issue.

McDermott argues against all three forms of extendibility considering both the extendibility of hardware and software. With Schmidhuber he holds that direct programming may not be extendible.[4] McDermott also holds that the extendibility of hardware is not guaranteed because of the lack of a smooth manifold as breakthroughs in hardware design are discontinuous and unpredictable.

Among the routes to extendibility Chalmers also considers brains embedded in a rapidly improving environment that result in an extended mind (*à la* Clark and Chalmers) similar to the scenario considered by Helighen. The section also considers extendibility and brain enhancement, something that will be elaborated on by Ray Kurzweil in the second volume.

The third premise, the amplification premise, claims that:

> Premise 3: If there is AI+, there will be AI++ (soon after, absent defeaters)

The premise relies crucially on assuming that increases in intelligence always lead to proportionate increases in the capacity to design intelligent systems. AI researcher Igor Aleksander argues that designing AI that can design machines as well as itself is much harder than Chalmers imagines and that increases in intelligence may lead to diminishing returns in design capacity. He holds that we will not be able to design machines that design machines as well as us in the foreseeable future.

Frank Tipler, the mathematical physicist and cosmologist (*The Anthropic Principle*), gives an alternative argument for the singularity, based on considerations from physics. Tipler argues that the entropy in a contracting universe cannot grow indefinitely and that the needed entropic cooling can only be supplied by an intelligence explosion. On Tipler's view, biological life forms will not survive the heat and pressure generated by a contracting universe and the only way to prevent an entropy explosion is for biological life forms to be either be uploaded or to design more robust AI that will be able to survive these extreme conditions. This means that the inevitability of the singularity is a direct outcome of our natural laws.

The second major part of Chalmers' article, 'Negotiating the Singularity', is concerned with maximizing the expected value of a post-singularity world.

[4] McDermott also holds that artificial evolution is not extendible, its interesting to compare some of his arguments with those of Shulman and Bostrom.

> In the near term, the question that matters is: how (if at all) should we go about designing AI, in order to maximize the expected value of the resulting outcome? Are there some policies or strategies that we might adopt? In particular, are there certain constraints on design of AI and AI+ that we might impose, in order to increase the chances of a good outcome?

Here Chalmers divides these constraints into external and internal constraints.

Section 6, 'Internal Constraints: Constraining Values', analyses ways in which we can maximize a positive outcome, for us humans, by designing AI with the right kinds of values. Chalmers distinguishes Humean approaches to AI, on which values are largely independent of intelligence (being built into a fixed utility function, for example), from Kantian approaches on which values are themselves rationally revisable.

Schmidhuber's Gödel machines rewrite their value functions and are Kantianin the sense of connecting morality and rationality even if they decide at some stage to rid the planet of sentient biological systems. Tipler's view also has a Kantian element in that he holds that an intelligence explosion must be based on honest agents and that if AI+ is to produce good science it must be honest. While in Schmidhuber's case constraining the evolving value system of his self-referential machines will significantly diminish their capacity Tipler's insistence on scientific honesty can only improve AI+ and AI++. However most AI researchers (and Chalmers) are more inclined to the Humean view that separates values and rationality.

In the second volume philosopher Barbara McCorduck holds that the human value system is too heterogeneous to lend itself to simplistic internal constraint scenarios. Like McDermott, and unlike AI researcher Murray Shanahan who entertains motivational defeater scenarios, McCorduck believes that structural defeaters are more likely.

AI researcher and mathematician Ben Goertzel who feels that the design of AI and AI+ must be constrained both internally and externally proposes an original solution:

> … the deliberate human creation of an 'AI Nanny' with mildly superhuman intelligence and surveillance powers, designed either to forestall Singularity eternally, or to delay the Singularity until humanity more fully understands how to execute a Singularity in a positive way. It is suggested that as technology progresses, humanity may find the creation of an AI Nanny desirable as a means of protecting against the

destructive potential of various advanced technologies such as AI, nanotechnology and synthetic biology.

Section 7 titled 'External Constraints: The Leakproof Singularity' explores ways of externally constraining the AI designs that might lead towards a singularity, especially constraining such AI to a virtual world from which it cannot leak into the real world.

AI researcher Roman Yampolskiy's article, 'Leakproofing the Singularity: Artificial Intelligence Confinement Problem', provides us with a detailed and well-reasoned analysis of this possibility.

Another external type of constraint mitigating unwanted outcomes is Robin Hanson suggestion to create legally binding contracts that AI+, for example, will have to abide by, minimizing intergenerational conflicts and guaranteeing our continued existence. In this scenario AI ++ will be legally obligated to upload us.

Francis Heylighen who advances an embedded approach rejects 'brain in a vet' scenarios and holds that confining AI to a virtual environment will result in greatly diminished capacity. Heylighen also holds that our sensory capacities honed by hundreds of millions of years of evolution cannot be successfully simulated unlike AI researcher Burt Voorhees who in the second volume explores the consequences of exponential advances in artificial sensory capacity.

In his article, 'Can Intelligence Explode?' AI researcher Marcus Hutter, who believes that the singularity is near, explores what it means to be inside and outside a singularity whose default state consists of interacting super-intelligent systems in a virtual world. Hutter believes that some aspects of this singularitarian society might be theoretically studied with current scientific tools (for example, superintelligent machine sociology) and that entering a singularity might be similar to crossing the event horizon of a black hole where we don't know that we have entered a singularity. However unlike crossing a black hole event horizon it is the outside which slows down to a crawl. Another reason that an outsider may miss the singularity altogether is that maximally compressed information is indistinguishable from random noise[5]. Arguing for a speed explosion, Hutter holds that what is meant by an intelligence explosion needs to be clarified by a better definition of universal intelligence.

In the second volume psychiatrist Chris Nunn holds that improving the definition of intelligence is complicated by the intrinsically contextual nature of information.

[5] In line with John Smart's transcention scenario.

The last major part of Chalmers article concerns uploading and the questions that it raises about consciousness and identity. We are introduced to destructive uploading as in 'serial sectioning', gradual uploading as in 'nano-transfer' and reconstructive uploading as a virtual resurrection. Will we survive uploading? Chalmers holds that the most agreeable form of uploading is probably gradual uploading in conjunction with a 'continuity of consciousness' approach to identity.
In his paper 'On Singularities and Simulations' philosopher Barry Dainton, who like Chalmers believes that the singularity scenario is certainly not out of the question, explores the mechanics of uploading and its relationship to identity. Much of his paper is devoted to an analysis of the possibility that we are already uploaded inhabitants in a virtual world, concluding that such a possibility may be quite higher than it seems. Dainton bases his argument on his take on Bostrom's 'simulation argument'. His simulation based approach towards distinguishing 'cartesian scepticism' from 'simulation scepticism' is another example of the relevance of the singularity scenario to some of our deepest philosophical questions about the nature of identity, reality and intentionality.

In the second volume Jesse Prinz holds that either we already are uploads and the singularity is here or we are not uploads and the singularity will not materialize arguing that in both cases we are doomed but adding that we have nothing to worry about.

Section nine, 'Uploading and Consciousness', asks whether an upload can be conscious. Chalmers holds a 'further fact' view of consciousness that leaves the question wide open. He suggests that an analysis of the gradual uploading scenario tends to support the functionalist approach.

Here philosopher Dan Dennett sets aside issues about the singularity and discusses Chalmers' 'further fact' view of consciousness. Dennett suggests that Chalmers' own 1996 work concerning gradual replacement shows that the 'further fact' view is unfounded, and offers some speculation about why Chalmers himself holds the view. The nature of this ongoing disagreement itself raises some interesting questions about the nature of philosophical truth and the philosophical endeavor. Ray Kurzweil also discusses issues about consciousness in his contribution to the second special issue.

Philosopher Richard Brown argues against the principle of organizational invariance and holds that uploading may force us to modify the conclusion of Chalmers' conceivability argument.

Section ten, 'Uploading and Personal Identity', asks whether uploading preserves our identity. In a comprehensive analysis of the

questions that relate identity, survival and uploading, Chalmers gives an argument based on destructive uploading that supports a pessimistic view and an argument based on gradual uploading that supports an optimistic view. While these arguments lead to diametrically opposed conclusions and cannot both be right we are not sure which view is correct. Chalmers reaches the conclusion that while holding a further fact view on consciousness is justified holding a 'further fact' view on identity is probably not.[6]

In her short paper 'She Won't Be Me', psychologist and memeticist Susan Blackmore, who is also sympathetic to the 'singularity soon' scenario, explains why contrary to (her take on) a pessimistic approach to the deflationary position, *that we never survive from moment to moment, or from day to day*, she finds this position to be exhilarating and liberating. Dainton also discusses issues about personal identity, holding that a continuity of consciousness approach to identity can resolve some of the problems encountered by the more orthodox 'Parfitian view'.

In the second volume philosopher Susan Schneider will also explore the way in which the very idea of the singularity forces us to reconsider identity especially due to enhancement.

To borrow from Plotnitsky, *the debate concerning the possibility of artificial intelligence goes back at least to Descartes and is, thus, coextensive with the history of modern philosophy*. As this symposium on the Singularity shows, and as this collection of responses shows, this debate is gaining a sense of urgency.

Acknowledgments

I enjoyed being the guest editor of this edition. I would like to thank David Chalmers for contributing the target paper, responding to the authors, and helping in many ways. I was privileged to collaborate on another JCS issue with Anthony Freeman. I am sad to see him go but he had a great run. I would also like to thank Ben Goertzel for sound advice, Arkady Plotnitsky, Bernard Baars, Hava Siegelmann and Yotam Hoffman for useful discussions, and Minerva San Juan, Ron Chrisley and TrinityDC University for their support.

[6] Joe Levin separates the hard problem into the problem of phenomenal content and the 'puzzle' of subjectivity. Perhaps it's possible to engage the latter while suspending the former.