The Unavoidable Charm of the Superintelligence and Its Risk

Gobbo, F.

Published in:
APA Newsletter on Philosophy and Computers

Moore's law can continue. It is true that Moore's law is occasionally invoked as an additional reason for why AI might become dangerous, but major works in the field do not assume that it would necessarily continue. *Intelligence Explosion – Evidence and Import* explicitly notes that it does not assume "the continuation of Moore's Law, nor that hardware trajectories determine software progress, nor that faster computer speeds necessarily imply faster 'thought' [. . .] nor indeed that AI progress will accelerate rather than decelerate." When *Superintelligence* mentions Moore's Law, it notes that "one cannot bank on this rate of improvement continuing up to the development of human-level machine intelligence." Finally, *Responses to Catastrophic AGI Risk* does not mention Moore's Law at all, other than to note that its continuation "depends on the existence of a small number of expensive and centralized chip factories, making them easy targets for regulation."

Finally, Floridi suggests that the main risk is not the appearance of superintelligence, but the misuse of more conventional digital technologies. While I disagree with him on the need to worry about superintelligence, I agree with him on conventional digital technologies certainly posing their own dangers as well. Work on avoiding the risks from superintelligence and more conventional technologies need not be mutually exclusive. There is currently only a very small number of people working full time on the risks from superintelligence, far fewer than there are people working full time on other risks such as pandemics. Effort put into protecting humanity from pandemics has not prevented other people from working on various issues of the digital era. Similarly, work focused on the implications of advanced AI can proceed without impacting the work done on other worthy causes.

NOTES

1. Luciano Floridi, "Singularitarians, Altheists, and Why the Problem with Artificial Intelligence is H.A.L. (Humanity At Large), not HAL," *APA Newsletter on Philosophy and Computers* 14, no. 2 (2015): 8–11.
2. Ibid., 8.

BIBLIOGRAPHY

Armstrong, Stuart, and Kaj Sotala. "How We're Predicting AI – or Failing To." In *Beyond AI: Artificial Dreams*, Pilsen, November 5-6, 2012, 52–75. Pilsen: University of West Bohemia, 2012.

Bostrom, Nick. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive, and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2. International Institute of Advanced Studies in Systems Research and Cybernetics, 2003.

Bostrom, Nick. "Existential Risks." *Journal of Evolution and Technology* 9, no. 1 (2002).

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Italy: Oxford University Press, 2014.

Branwen, Gwern. "Slowing Moore's Law: How It Could Happen." 2012/2015. http://www.gwern.net/Slowing%20Moore%27s%20Law

Chalmers, David. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17, no. 9-10 (2010): 7–65.

Floridi, Luciano. "Singularitarians, Altheists, and Why the Problem with Artificial Intelligence is H.A.L. (Humanity At Large), not HAL." *APA Newsletter on Philosophy and Computers* 14, no. 2 (2015): 8–11.

Muehlhauser, Luke, and Anna Salamon. "Intelligence Explosion: Evidence and Import." In *Singularity Hypotheses*, 15–42. Springer Berlin Heidelberg, 2012.

Müller, Vincent and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, edited by Vincent Müller. Berlin: Springer, 2014.

Russell, Stuart and Peter Norvig. *Artificial Intelligence: A Modern Approach*, 3rd edition. Pearson, 2009.

Sotala, Kaj. "Advantages of Artificial Intelligences, Uploads, and Digital Minds." *International Journal of Machine Consciousness* 4, no. 01 (2012): 275–91.

Sotala, Kaj, and Roman V. Yampolskiy. "Responses to Catastrophic AGI Risk: A Survey." *Physica Scripta* 90, no. 1 (2015): 018001.

Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*. Oxford: Oxford University Press, 2007.

The Unavoidable Charm of the Superintelligence, and Its Risk

Federico Gobbo

UNIVERSITY OF AMSTERDAM

Readers of the APA newsletter are used to speculation and theoretical debates, being philosophers. The last one is the fierce attack of Floridi and the defense of Sotala to the debate about the future of AI and the theoretical possibility of Singularity, Superintelligence, or AI+ (mainly Chalmers), according to the different authors. Is a truly autonomous, morally independent, (bio)mechanical being that can control our digital technologies against us plausible? In short, Floridi argues that it is theoretically possible, but so implausible that it is not worth spending a word on it—of course, he has to spend some words in order to say it, with is somehow paradoxical. And his text calls for reactions, as the advocates of singularity are treated as if they were members of a sect. Sotala adheres to the wording used in Bostrom's book, who—not by chance—uses the word "Superintelligence" instead of "Singularity."

I invite the reader to take a step backwards and look at this debate with more distance. Let us try to recall what we have learned from the history of ideas in AI. Unfortunately, the tradition of AI is sometimes forgotten in such debates because scholars are urged to quote recent papers and recent authors. We lose our past; we lose our memory. Floridi underlines the proximity between the Singularitarians and Hollywood. I want to extend his metaphor telling that, in my view, this debate is like a new movie with an old plot, like a reboot of a classic of science fiction. In the old days, the debate was about the plausibility of Good Old Fashioned Artificial Intelligence (GOFAI). I tried to read the main positions in this debate, but I failed to find something new. As in any good reboot, some details are different, but the core message is not. What is the concrete result of the debate about GOFAI? Essentially, AI has lost credits because of this speculation. The concrete, operative results of research came from the so called "weak AI," which, in short, rejects all the theoretical problems of true AI as uninteresting or pointless (as Floridi says), adopting an *a posteriori* perspective: an artificial agent which shows intelligent *behavior* can be considered intelligent, regardless if the *process* behind its behavior is really intelligent.

I argue that the point is that the risk we are facing now is a new discredit of AI. But (weak) AI is more and more present in our daily lives than before. That is why I signed the open letter published in 2015 within the charity Future of Life about the research priorities for "robust" (an internal feature with epistemological consequences) and "beneficial" (a moral concern, as it addresses humankind) artificial intelligence. And I can guarantee to the readers that I do not adhere to any church, Singularitarians and Atheists—to use Floridi's terms—included. Sotala mentions that letter as if the whole debate about the plausibility of GOFAI/Singularity were supported by that. Well, it is not. It suffices to quote the opening of the letter itself:

> Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents—systems that perceive and act in some environment. In this context, "intelligence" is related to statistical and economic notions of rationality—colloquially, the ability to make good decisions, plans, or inferences.

This definition of "intelligence" comes from the tradition of weak AI, and it *a priori* excludes the debate of GOFAI/Singularity as completely irrelevant. We desperately need moral philosophers collaborating with hard science researchers in order to achieve the goal of beneficial AI. Now. Possibly, short-termed. It is completely irrelevant the speculations of researchers in the field in the long-term, mentioned by Sotala: experience shows that even great minds playing with the game of futurology ultimately proved to be completely wrong. But there is a more urgent consideration to be made in this sense. As Keynes said, in the long run we are all dead. The risks we are facing are today, not tomorrow: a badly designed multi-agent system can be a disaster when applied to a large scale, interacting with human beings in an unpredicted manner.

I think that the main risk inside the Superintelligence is the risk of losing the focus on the real problems. But then, why are so many people worried? What is the explanation for it? I have my own opinion on that. The computational turn tremendously complexified our lives. We, human beings, fear complexity because we feel that we are losing our control on reality. The reaction is to look for a single reference point where all relevant causes can be addressed. And here it is: Superintelligence, an Orwellian Big Brother that controls everything. A *single* artificial mind. After all, many among us still did not learn the lesson of the Internet, which is *a network with no central point* that controls everything.

I invites all researchers, especially the younger, to devote their energies to the real problems of artificial intelligence in our contemporary world, letting speculation into the realm of science-fiction literature and Hollywood movies.

Some comments on Luciano Floridi's The Ethics of Information

Jacques Bus
SG DIGITAL ENLIGHTENMENT FORUM

Many of us ask ourselves how we have to understand and live in a world with an increasing number of autonomous technical information systems and a society that through digitization reaches levels of complexity that seem to make our democratic and ethical rules and institutions unfit for their tasks.

Luciano Floridi (LF) has done an impressive job addressing these problems in his book *The Ethics of Information*. His philosophical approach is, for me (mathematician of origin), refreshing in the sense that his thinking is built up in highly analytical terms. He explains mathematical concepts like "level of abstraction," "complex and self-emergent systems," and the concept of "entropy" from thermodynamics, which is also used (but differently) in classic information theory.

The introduction of the term "metaphysical entropy" and how this is used to define four ethical principles of Information Ethics (IE) did raise questions for me. As I was reading, I sometimes asked myself where this could lead. The problem of ethics in general does not particularly lend itself to a mathematical or quantifying approach. However, a clear and satisfying answer follows on page 315 as a response to some of the criticisms. Floridi states there:

> IE is equally reasonable: fighting the decaying of *Being* (metaphysical entropy) is the general approach to be followed, not an impossible and ridiculous struggle against thermodynamics, or the ultimate benchmark for any moral evaluation, as if human beings had to be treated as mere numbers.

So good, ethical behavior is fighting the decaying of being or, in LF's terms, fighting the decrease of metaphysical entropy in the overall system.

The concept of Global Information-Ethics, developed throughout the book, is not simply defined in a few sentences without risking wrong interpretations. The interested person will have to read to the end. An important aspect is that the actors in the ethical space are not restricted to human agents and patients, but include all information entities; hence, also non-intelligent objects and creatures, autonomous technical systems, organizations or communities, etc.

A second important issue is the proposal to develop a global informational ontology for a global digital world.

I am not an ethicist, nor a philosopher. Hence I cannot judge the book on its value for those scientific communities. My interests lie in the interaction between digitization and society (with the individuals living in it) and the policy consequences and requirements. What interests me most