The Unavoidable Charm of the Superintelligence and Its Risk

Gobbo, F.

Publication date
2016
Document Version
Final published version
Published in
APA Newsletter on Philosophy and Computers

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.