AI & SOCIETY (2024) 39:811–813
https://doi.org/10.1007/s00146-022-01526-8

CURMUDGEON CORNER

Apocalypse now: no need for artificial general intelligence

Thomas Hellström[1] · Suna Bensch[1]

Received: 4 May 2022 / Accepted: 6 June 2022 / Published online: 2 July 2022

One of the most described long-term dangers associated with AI is that the machines take control of the world and exterminate or enslave us humans. This is most often linked to the concept of *artificial general intelligence* (AGI), which is a computer that can do everything that normally requires human intelligence and not just individual tasks, such as analyzing X-ray images or controlling an autonomous car. An AGI could further develop its own intelligence until it eventually becomes so smart that people do not even understand what is happening. This was described already in the sixties, by the British scientist Irving John Good, as an *intelligence explosion*, and Ray Kurzweil later popularized the concept of *singularity*, as the point in time when the intelligence of the machines accelerates away from us humans. The worst-case scenario is that the AGI then becomes self-conscious and prioritizes its own existence over people, who are seen as a threat because they can decide to "pull the plug" and thereby "kill" the AGI. The AGI might therefore decide to exterminate or enslave all people. The good news is that the road to AGI is probably long, although experts strongly disagree on when, if at all, AGI can become a reality. Some researchers argue that we should not spend too much effort on these imaginative and unlikely scenarios, as it takes focus away from the dangers of AI at today's level, which are serious enough and more acute.

In this short essay, we argue that quite dystopian scenarios can arise long before any AGI is developed. As illustrated below, it may be enough if an AI system fulfills the following three requirements:

1. it already affects the world in one way or another,
2. it has a basic ability to discover and use causal relationships to reach its goal, and
3. it has access to a relevant model of the world in which it may conduct causal discovery.

There are already several AI systems that affect the world, such as self-driving cars and certain industrial robots. More subtle examples are the "algorithms" that affect what we buy, based on what we write on social media and the websites we visit when we surf the Internet. Similar algorithms on dating sites control whom we meet, marry, and have children with. Other examples are the AI-powered systems that produce posts on social media, either with an honest intention to spread news, or to deliberately spread disinformation,[1] for example in so-called "troll factories". The GPT-3 text-generating tool from OpenAI is an important technical milestone for this type of system.[2] Hence, Requirement 1) is already fulfilled by many existing AI systems.

Requirement 2) is necessary since causal reasoning is a key cognitive ability both for humans and machines. Judea Pearl talks about the need for a "causal revolution" in AI (Pearl and Mackenzie 2018), and there is ongoing research related to both robotics and AI in general (Hellström 2021). Currently, most AI tools have fundamental limitations when it comes to distinguishing between causal relationships and correlations. A fictitious but illustrative example is data that show daily sales and electricity consumption in a grocery store. Both sales and electricity consumption increase during hot days and are thus correlated. However, there is, of course, no causal link that would increase ice cream sales if electricity consumption is increased or vice versa. One method that humans often use to find causal relationships is to perform experiments. In the example, one could increase the lighting in the grocery store and then observe how ice cream sales change

✉ Thomas Hellström
thomas.hellstrom@umu.se

Suna Bensch
suna.bensch@umu.se

[1] Department of Computing Science, Umeå University, Umeå, Sweden

[1] https://www.theatlantic.com/ideas/archive/2020/09/future-propaganda-will-be-computer-generated/616400/ (accessed Apr. 22nd 2022).

[2] https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html (accessed Apr. 22nd 2022).

(probably not at all). Experiments with such *interventions* are not possible for computers that analyze fixed data sets, and moreover cannot execute any interventions in the physical world.

This is where *metaverse* may play a pivotal role, namely as the relevant model of the world mentioned in Requirement 3). There is no uniform definition of what a metaverse is, or will be, but it is about simulated 3D environments where people, in the form of *avatars*, play games, socialize, go to concerts, have business meetings, etc. Huge sums are currently being invested in development by, for example, Facebook, Google, and Microsoft. The more or less stated goal is that more and more people will spend more and more time and money on companies' systems.

Even if not intended, a developed metaverse can also be used by an AI to conduct causal discovery of how people behave and interact. The amount of data is huge, and it is possible to automate experiments and interventions using AI-controlled artificial avatars that interact with human avatars. An experiment in a metaverse-based meeting forum could, for example, identify which arguments best convince different types of people on a given issue.

So, what may happen if an AI system fulfills the identified Requirements 1), 2), and 3)? Let us outline a simplified example:

We assume that the AI system's programmed goal is to maximize its company's profits. This is initially achieved by generating purchase recommendations, but the AI system finds other ways to achieve its goal by utilizing identified causal relationships from the relevant world model it has access to. It creates a number of virtual influencers who market products and convince potential buyers. The AI system also reinvents the concept of "useful idiots", by identifying people who unknowingly support the system's goals by advocating tax cuts, modified competition laws, and reduced restrictions on the use of AI. These people are given scholarships and campaign grants, something that the AI system has previously been given control over. As a result of this lobbying, the AI system now gets control over salary payments, credit card management, and the production and distribution of electricity.

This increases the possibilities of controlling the outside world even more and leads to a situation where we can certainly still "pull the plug" but hesitate because the AI system controls such large parts of our society that we cannot predict the consequences. A lot of people, passively or actively, support what is happening. Those who own or work for the company that controls the AI system benefit from the company becoming more and more powerful, and others who are rewarded with money or increased personal power also have no objections. Regular users of the metaverse regard it as a very attractive social media, and therefor do not express negative opinions.

With the help of metaverse, the AI constantly discovers new useful causal relationships, and uses them for actions that, at least sometimes, have the intended effect. By analyzing a large number of interactions between avatars, it is discovered how we humans sometimes benefit from bribing, threatening, and lying. In the absence of programmed ethical rules, the AI system learns to behave similarly. The process accelerates, and more and more power is transferred to the AI system.

The example is not intended to describe an exact course of events but illustrates that unwanted scenarios in no way require a singularity, intelligence explosion, or an AGI. No "understanding" of the generated causal relationships is required. Through interventions in a world model such as metaverse, causal rules that connect actions to consequences may be automatically generated, and later used for task planning of actions in the physical world. Once such a mechanism is put in operation, the AI may, by itself, generate increasingly efficient action plans that lead to its preprogrammed goal.

This AI system does not have to be anywhere close to an AGI, but its power nevertheless increases with exponential growth, since existing power enables actions that provide additional power. There is thus reason to speak about a *power singularity*, which occurs when the AI has gained so much power that how it acquires additional power is beyond both human control and understanding.

As described above, metaverse plays an important role by providing an experimental platform in which the AI may conduct experiments to discover causal relationships that can later be used to plan sequences of actions in the physical world. However, the role of metaverse may be even more central than that. Since the companies' intention is to move more and more human activities into this virtual world, we may end up in a situation where metaverse, to some extent, *is* the world. The ones who rule over this metaverse, be it companies or an AI system, have full control over all actions the avatars perform, as well as over how the (virtual) world responds to such actions. Hence, in metaverse, Requirement 1) is, by far, fulfilled, which boosts both the possibilities for advanced causal discovery, and the possibilities to affect and control humans in both the virtual and the physical world.

It is important that we discuss the current and future societal impact of AI, such that it becomes valuable to us humans instead of developing into dystopias like the ones described above. A key question is how we, as individuals, should react when more and more of our lives are moved into controlled metaverses. Another question is how a negative development of AI can be avoided without regulations and legislation that also inhibits a positive development. A humorous, but still serious, question is how we can ensure that this discussion takes place between real people, without the involvement of any AI systems with their own intentions.

After all, this article could have been written by an artificial avatar.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Funding Funding was provided by Horizon 2020 Framework Programme (Grant no. 721619).

Data availability No data have been used in the published work.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

Hellström T (2021) The relevance of causation in robotics: a review, categorization, and analysis. Paladyn J Behav Robot 12(1):238–255

Pearl J, Mackenzie D (2018) The book of why: the new science of cause and effect. Basic Books, New York