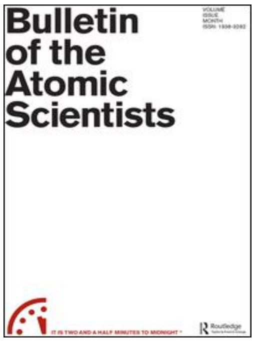

Bulletin
of the
Atomic
Scientists

Bulletin of the Atomic Scientists

ISSN: 0096-3402 (Print) 1938-3282 (Online) Journal homepage: https://www.tandfonline.com/loi/rbul20

The possibility and risks of artificial general intelligence

Phil Torres

To cite this article: Phil Torres (2019): The possibility and risks of artificial general intelligence, Bulletin of the Atomic Scientists, DOI: 10.1080/00963402.2019.1604873

To link to this article: https://doi.org/10.1080/00963402.2019.1604873

Published online: 26 Apr 2019.

BULLETIN OF THE ATOMIC SCIENTISTS
https://doi.org/10.1080/00963402.2019.1604873

Routledge
Taylor & Francis Group

The possibility and risks of artificial general intelligence

Phil Torres

ABSTRACT
This article offers a survey of why artificial general intelligence (AGI) could pose an unprecedented threat to human survival on Earth. If we fail to get the "control problem" right before the first AGI is created, the default outcome could be total human annihilation. It follows that since an AI arms race would almost certainly compromise safety precautions during the AGI research and development phase, an arms race could prove fatal not just to states but for the entire human species. In a phrase, an AI arms race would be profoundly foolish. It could compromise the entire future of humanity.

KEYWORDS
Arms races; artificial intelligence; international relations; superintelligence

From programs that can recognize faces to those that can understand human speech, artificial intelligence developers have made huge strides in designing AI systems capable of analyzing and interpreting data in meaningful ways. But these technologies, useful as they may be, represent narrow forms of intelligence. Researchers hope to expand the capabilities of AI by creating systems with the ability to explain their results, reason abstractly, and learn in a less data-intensive and more human-like way. While the US military's Defense Advanced Research Programs Agency is investing in these sorts of efforts, some researchers see an even greater potential for revolutionizing the cognitive abilities of machines: The ultimate dream of AI research is to develop so-called artificial general intelligence, or AGI.

The fact is that an AGI could be the most powerful technology ever invented. Theorists have wildly divergent views on when an AGI breakout might occur, with some speculating that it won't happen for many decades to come. If and when an AGI breakout happens, it could bring about a revolution far more transformative in human history than the Neolithic and industrial revolutions by being, as I.J. Good famously wrote, the "last invention that [humans] need ever make" (Good 1966).

Whereas past technologies have been tools used by humans, AGI would constitute an agent in its own right. Even more, a system with the same capacities as the human mind would automatically constitute a quantitative superintelligence by virtue of its ability to process information at least a million times faster than the human brain. If students in the United States spend an average of 8.2 years earning a PhD, an AGI could accomplish this in only a few minutes. In fact, this is probably an overestimate of the time required, since AGIs would have direct access to a variety of systems that are narrowly superintelligent, such as calculators, as well as online knowledge databases like Wikipedia.

AGI could be just a stepping stone

Furthermore, AGI could be a momentary flash between sub-human-level AI and artificial superintelligence. An artificial superintelligence is a system that could significantly outperform every possible human in all cognitive domains. The reason is that whatever the AGI's programmed goals might be, enhancing its cognition by tinkering with its own source code would constitute an intermediary step toward achieving these goals. Thus, whether the AGI wants to calculate as many digits of the square root of two as possible, find a cure for cancer, build a few cobalt bombs, or just design a new suburban neighborhood, the AGI has an interest in becoming smarter.

It follows that, insofar as the AGI acquires what NYU professor David Chalmers refers to as an "extendible method," or one that can be easily improved, boosting its own intelligence would be among its very first actions (Chalmers 2010). The AGI could thus initiate a positive feedback loop whereby each augmentation enables it to further augment itself, resulting in an intelligence explosion. In a matter of minutes, hours, or days, humanity could be joined by a being not merely more intelligent than humans, but vastly more capable of solving problems in a wide range of cognitive domains.

Researchers are working on AGI now

At present, the field of AGI research and development is marked by overlapping goals and considerable

CONTACT Phil Torres ✉ philosophytorres@gmail.com

cooperation, with some projects being run by the same people. But that could change. In fact, several researchers are concerned about an AGI arms race involving autonomous superintelligent machines in the coming years or decades. Many aspects of a potential AGI technology suggest that developers would face a great challenge controlling it. Governments may have a huge incentive to be the first to develop AGI. The country that crosses the finish line first may have a decisive advantage that not even nuclear weapons can provide.

Right now, there are numerous projects around the world working toward creating an AGI. Co-founder of the Global Catastrophic Risk Institute Seth Baum lists 45 such endeavors, including DeepMind, OpenAI, GoodAI, CommAI, CogPrime, SingularityNET, and the Human Brain Project (Baum 2017). Twenty of these are affiliated with academic institutions, the rest being connected to private corporations, public corporations, nonprofits, and governments. Furthermore, a clear majority – 40 to be exact – have what Baum refers to as humanitarian or intellectualist goals. That is to say, they are pursuing AGI for the express purpose of benefiting humanity as a whole and to solve intellectual problems in science and mathematics.

Not many of these projects have connections with the military, and the eight that do are primarily US-based academic projects that receive military funding. Only one project, DSO-CA, is based in a military-defense institution, namely, Singapore's DSO National Laboratories, which conducts national defense research. However, Baum reports that the military connections of 32 of the 45 AGI efforts are unspecified. Since Baum's survey relies on publicly available data, there could be any number of covert AGI projects being run by governments or corporations – or perhaps even small groups or single individuals.

The possibility of doom

All AGI developers will have to contend with the control problem – i.e. the challenge of creating algorithms that are more generally intelligent than humans without losing control of those algorithms. (Note: algorithms form the mathematical basis of AI programming.) This may not sound especially difficult, but a closer look suggests that it could constitute perhaps the most formidable intellectual conundrum that humanity has ever had to overcome. Furthermore, there are reasons for believing that the default outcome of failing to solve the control problem will be severe – not just for one or two states or a subset of the human population, but for humanity as a whole.

Even more, given the possibility – or probability – of a fast takeoff, whoever creates the first AGI will forever push humanity past a Rubicon and into a fundamentally new era. Could states or other AI developers hit the pause button on a failed AGI project?

There are plenty of technologies that one can lose control of without devastating, or bad, or merely non-good consequences. But AGI is different. Since it would by definition rival or far exceed human problem-solving capacities, its ability to pursue whatever goals it is given could transcend a developer's ability to stop it if unintended deleterious effects occur. Furthermore, it doesn't seem to be the case that a superintelligent algorithm would automatically realize that causing harm in pursuit of its goals is bad and should be corrected.

The control problem becomes clear upon considering several thorny philosophical concepts. Together they yield a conundrum that might not be solvable:

> ***The orthogonality thesis.*** Some theorists believe a wide range of final goals can be combined with a wide range of intelligence levels. Thus, there's nothing incoherent about a superintelligent machine caring only about playing tic-tac-toe for the next billion years. Nor is there any principled reason to think that a superintelligent machine whose goal system is programmed to count the blades of grass on Harvard's campus would stop and think, "I could be using my vast abilities to marvel at the cosmos, construct a 'theory of everything,' and solve global poverty. This is a silly goal, so I'm going to refuse to do it." Algorithms do what they're told.
>
> ***The instrumental convergence thesis.*** An AGI would likely pursue its goals relentlessly, no matter where they ranked in the hierarchy of human values. Toward that end, it would work toward several intermediary goals like cognitive enhancement. Other such goals could include acquiring goal-related resources, maintaining the integrity of its system of final goals, and preventing humans from shutting it down. Therein lies one hypothetical avenue for a dangerous AGI to emerge. In an interview with HuffPost, Oxford University professor Nick Bostrom gave the example of an AI designed to make as many paper clips as possible: "The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans." The danger here is that the human species as well as the biosphere and planet are resources that could be used to achieve the AGI's final goals, whatever they are. (See Häggström 2018 for further discussion.)

> ***The complexity of value thesis.*** Could AGI developers simply program in human values? First, these values have what computer scientists call a high Kolmogorov complexity, meaning that they cannot be compressed in a simple rule. For example, consider the numbers 123123123 and 352695142. The first has a lower Kolmogorov complexity because it can be compressed into 123 three times, whereas the second number cannot be described so simply. Human values resemble the second number much more than the first. Even if developers knew what human goals are, there's still the problem of getting them inside the AGI.
>
> ***The perplexity of value thesis.*** What even are human values or goals? Many philosophers throughout history have debated fundamental issues relating to normativity, or the question of what ought to be the case aside from what is the case. Rather than converge on a single or small set of views, as has occurred in the sciences, one finds the opposite: a proliferation of different perspectives, each with their own prominent advocates. A survey from 2014 found widespread disagreement among professional philosophers about normative ethics, moral motivation, and meta-ethics, among other topics.
>
> The situation is equally hopeless within the domains of economics and politics; there are massive disagreements about whether democracy is better than authoritarianism, capitalism is better than communism, and so on. Even among those who pick one ideological side over another, there are nontrivial divergences of opinion about an interminable number of details. Thus, potential AGI developers must give a lot more thought to what sorts of commitments could be termed human values.
>
> ***The fragility of value thesis.*** It could also be that value systems have all-or-nothing properties: either they work perfectly or fail miserably. And it might be that if a single component value is missing, then one gets a radically different outcome than what one wanted. The classic analogy is that dialing nine out of 10 phone number digits correctly doesn't get someone who's 90 percent similar to the intended receiver. It might not scale linearly like this, meaning that a goal system that needs 1,000 parts to function properly but only has 999 of these parts in place when it attains AGI status – and subsequently ignites an intelligence explosion – could have cataclysmically bad consequence as it proceeds toward its goal.

Then there's the so-called relative speed thesis – the idea that the electrical potentials in computer hardware can process information much faster than the action potentials in human brains. Even a brain that's emulated on a computer would immediately attain superintelligence in some domains like calculation through access to online information. To it, the outside world would be virtually frozen in time. There's also the rapid capability gain thesis, which describes how an AGI would quickly work to improve its own intelligence. Considering these theses together, one can grasp how combustible an AGI might be.

If AGI developers don't get things exactly right on the first go, an AGI that's programmed to, say, cure cancer might quickly destroy the global ecosystem by converting it into research laboratories, thus bringing about humanity's demise. This may sound absurd, but ultimately the lesson is that even after extensive reflection on how things could possibly go wrong, developers might still be missing something critical – something that causes the AGI to wreak existential havoc in pursuit of its programmed goals. To paraphrase the late Stephen Hawking, if advanced artificial intelligence isn't the best thing to happen to humanity, it will almost certainly be the worst.

AGI arms races

An AGI arms race could be extremely dangerous, perhaps far more dangerous than any previous arms race, including the one that lasted from 1947 to 1991. The Cold War race was kept in check by the logic of mutually-assured destruction, whereby preemptive first strikes would be met with a retaliatory strike that would leave the first state as wounded as its rival. In an AGI arms race, however, if the AGI's goal system is aligned with the interests of a particular state, the result could be a winner-take-all scenario.

Yet states may be tempted to pursue AGI with growing urgency because whoever creates the first AGI could gain total control over civilization forever. Russian President Vladimir Putin made this point when he declared in 2017: "Artificial intelligence is the future, not only for Russia but for all humankind … Whoever becomes the leader in this sphere will become the ruler of the world." Although it's unclear whether Putin was thinking about AGI specifically or AI systems generally, he's probably right.

This situation could be exacerbated by the fact that, whereas the Cold War involved a bipolar configuration of just two actors engaged in arms racing, there could be far more than two adversarial actors attempting to create AGI. According to a 2016 game-theoretic analysis by Oxford University researchers, the risk of a dangerous AGI arms race rises when the number of AGI projects increases, information is freely shared among groups, and groups harbor enmity toward each other. There are many AGI research and development projects, and many make their code open-source. Fortunately, the third condition is not currently the case, although it would be in an arms race scenario.

The dangers of AI denialism

The critical importance of getting states and other potential competitors to take seriously the control problem is why the phenomenon of AI denialism is so dangerous. This term refers to dismissals of the AGI threat that distort and misrepresent the concerns of AI safety experts.

In the best-selling book *Enlightenment Now*, Harvard University psychologist Steven Pinker compared worrying about superintelligent computers to “a 21st-century version of the Y2K bug.” AI disasters, he wrote, are “sometimes called the Robopocalypse.” In fact, that is Pinker’s own misleading term – most people that think seriously about AGI don’t believe that Terminator-style robots will rise up and dominate humanity. Those who worry about AGI are not at all engaging in anything like the hyperbolic panic that surrounded the Y2K bug. But in a 2015 article, Pinker dismissed them:

> AI dystopias project a parochial alpha-male psychology onto the concept of intelligence. They assume that superhumanly intelligent robots would develop goals like deposing their masters or taking over the world. But intelligence is the ability to deploy novel means to attain a goal; the goals are extraneous to the intelligence itself … It’s telling that many of our techno-prophets don’t entertain the possibility that artificial intelligence will naturally develop along female lines: fully capable of solving problems, but with no desire to annihilate innocents or dominate the civilization.

Pinker thus rejects an entire body of work in AGI safety with the flick of a wrist. AGI could have irreversible, world-transforming consequences. It is therefore imperative moving forward that all parties involved in the creation of AGI recognize the enormity of getting AGI wrong. In the absence of cooperation, and especially if projects cut safety corners in an effort to reach the AGI finish line first, all of humanity could suffer terribly.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes on contributor

Phil Torres is an author and scholar whose work focuses on existential risks to civilization and humanity. His most recent book is *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Currently a visiting researcher at the Centre for the Study of Existential Risk at Cambridge University, he’s completing what will be his fourth book, titled *A Brief History of Human Extinction*.

References

Baum, S. 2017. “A Survey of Artificial General Intelligence Projects for Ethics, Risk and Policy.” Global Catastrophic Risk Institute Working Paper 17-1. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741

Chalmers, D. 2010. “The Singularity: A Philosophical Analysis.” *Journal of Consciousness Studies* 17 (9–10): 7–65.

Good, I. J. 1966. “Speculations Concerning the First Ultraintelligent Machine.” *Advances in Computers* 6: 31–88.

Häggström, O. 2018. “Challenges to the Omohundro-Bostrom Framework for AI Motivations.” *Foresight*. doi:10.1108/FS-04-2018-0039.