Title
Artificial General Intelligence and the Future of the Human Race

Journal
Berkeley Scientific Journal, 16(2)

ISSN
1097-0967

Author
Pavlacka, Bryon

Publication Date
2012

DOI
10.5070/BS3162016103

Undergraduate

Artificial General Intelligence and the Future of the Human Race

Bryon Pavlacka

Artificial Intelligence is all around us. It manages your investments, makes the subway run on time, diagnoses medical conditions, searches the internet, solves enormous systems of equations, and beats human players at chess and Jeopardy. However, this "narrow AI," designed for solving specific, narrow problems, is something distinctly different from Artificial General Intelligence, or "AGI", true thinking machines with human-like general intelligence (Wang, Goertzel, & Franklin, 2008, p. v). While AGI is not rigidly defined, it is often envisioned as being self-aware and capable of complex thought, and has been a staple of science fiction, appearing prominently in popular films such as 2001: A Space Odyssey, Terminator, and I, Robot. In each of these films, the machines go beyond their original programmed purpose and become violent threats to humans. This is a possibility which has been pondered extensively by those working in the field of AGI research. Thinkers like Ray Kurzweil, Ben Goertzel, and Hugo De Garis think that we are entering into a world of extremely intelligent machines (Kurzweil 2005; Goertzel & Pennachin, 2007; De Garis, 2008). This article will discuss some of the ideas that researchers have on how AGI relates to the wellbeing of humans, including how the machines can help us and how they could potentially harm us.

> ***"Narrow AI is already used by the militaries of first world countries for war purposes."***

One scenario in which generally intelligent machines go bad and become a threat is if they end up in bad hands. Such machines, in the hands of small, politically motivated terrorist groups or large military organizations, could be used as weapons. AGI could offer such groups the ability to spy, gather, and synthesize information, as well as strategize attacks against the rest of the population. Developers of AGI will have little knowledge of whose hands their technology will end up in; they could unknowingly be constructing deadly weapons to be used against humanity. Of course, such threats are not imaginary future possibilities. Narrow AI is already used by the militaries of first world countries for war purposes. Consider drones such as the Northrop Grumman X-47B, an Unmanned Combat Aerial Vehicle that is being tested by the US Navy (DefenseTech.org, 2011). That's right, there is no pilot. Of course, the drone can be given orders, but the exact way in which those orders are carried out will be left up to the drone's Narrow AI system. Whether such systems will ever be extended toward general intelligence is currently unknown. However, the US military has shown interest in producing and controlling generally intelligent killing machines as well, as made evident by a paper called "Governing Lethal Behavior" by Ronald C. Arkin. The paper was commissioned by the U.S. Army Research Office and provides theory and formalisms for the implementation of a lethal AGI machine (Arkin, 2008). The paper describes a way in which a machine can be restricted to "ethical" behavior determined by the creator. The author optimistically hopes that his proposed formalisms may lead to generally intelligent battle drones that are more ethical in battle than humans are, yet the ability to define "ethical actions" remains the privilege of the machines' engineers (Arkin, 2008, p.62). Due to the potential for AGI to be used as a weapon, the production of such machines carries many of the same moral ramifications as the production of other weapons.

Another threat to humanity is the possibility that a good machine, one specifically created to be benevolent, may go bad, as was the case with Hal 9000 in 2001: A Space Odyssey. Evolutionary and learning algorithms may lead to a system that is essentially a black box, something so complicated that experts may be unable to understand its inner workings completely. As with humans, such machines may have extremely complex psychologies, so the potential for malevolence is non-zero (Goertzel & Pennachin, 2007). Even if special constraints are placed on the behavior of such systems, rules like "do not kill" could potentially be overwritten after successive updates initiated by the AGI system itself (Singularity Institute for Artificial Intelligence [SIAI], 2001, para. 2).

Such a scenario may become greatly feared by the public, leading to what one researcher calls "The Artilect War." In Hugo De Garis' essay, "The Artilect War:

A Bitter Controversy Concerning Whether Humanity Should Build Godlike Massively Intelligent Machines," he proposes that humans will gain the capacity to construct massively intelligent machines called Artilects and that humans will fight over whether or not to construct them. He postulates the emergence of three major camps:

> *"As with humans, such machines may have extremely complex psychologies, so the potential for malevolence is non-zero."*

"Cosmists" who want to build artilects, "Cyborgs" who want to become artilects, and "Terrans" who are opposed to the construction of artilects (De Garis, 2008). De Garis outlines the social climate that will lead to the Artilect War. "The 'species dominance' debate has already started," De Garis argues. "The fundamental question is whether humanity should build artilects or not. The issue will dominate our global politics this century, and may lead to a major war killing billions of people" (De Garis, 2008, p. 440). De Garis' argument is understandable, though not completely convincing. Kurzweil argues that there won't be many Terrans. Instead, he claims that just about everyone will be using technology to improve their own abilities, that there will be no fine distinction between machines and us. The other possibility that De Gargis neglects is that the artilects may be created before would-be-terrans are even aware of the possibility of their construction. The specificity of De Gargis' predictions seems too narrow and the array of alternative possibilities is not sufficiently explored in his essay.

If AGI is developed, there are many benefits which would likely follow: it could lead to exponential advances in every scientific field. Grad students would no longer have to be worked to death since machines could do much of the epistemic labor. AGI could also be applied to fields which require an extraordinary amount of training. Doctors and other professionals could be replaced by efficient machines that don't get tired, don't require extensive training, and make fewer mistakes. In the future, access to cheap medical care could become the norm. One of the most optimistic futurists of our time, Ray Kurzweil, believes that exponential advancements in technology will lead to a technological Singularity, an intelligence explosion likely owed to the development of something called Seed AI. The term Seed AI is used for any intelligent system that is capable of very rapid exponential gains in its intelligence. Seed AI is imagined to be capable of modifying its own programming to construct a smarter self; this updated version would then be even better at programming, thus being capable of creating even smarter subsequent updates, and so on (SIAI, 2001, para. 2). This leads to an infinite feedback loop of intelligence. But will the result be the god-like Artilects which haunt the dreams of many science fiction writers or will it be the hoped-for Singularity that Kurzweil predicts?

Kurzweil has a very good track record when it comes to predicting future technological trends. When Kurzweil published The Age of Intelligent Machines in 1990, he accurately predicted that computers would outmatch the world's greatest human chess player by 1998, which happened in 1997. He also predicted the rise in popularity of the internet as well as cell phones (Kurzweil, 1990). In his more recent book, The Singularity is Near, Kurzweil predicts that by 2020, PCs will have the same raw processing power as a human brain, by 2030, humans will be able to upload their minds onto computers, and by 2045, the Singularity will emerge as computers become smarter and more capable than humans (2005). While Kurzweil acknowledges that there is much uncertainty about the details, he embraces the possibility of a Singularity and argues that it offers us many benefits and will likely be friendly. Kursweil also claims that we have gradually been merging with machines and that this process will continue. In a video interview with Big Think, Kursweil states:

> *"Just about everyone will be using technology to improve their own abilities, that there will be no fine distinction between machines and us."*

If you talk to a biological human, they will have lots of non-biological processes going on in their body and brain. Those computers will be out on the clouds, and the thinking of that "person" isn't even just in their body and brain even in the non-biological portion, it's out on the cloud. So it's going to be all mixed up, there's not going to be a clear distinction between human and machine. (Big Think, 2011)

Kurzweil believes that the vast majority of humans will gladly merge with the machines and become cyborgs. Of course, Kurzweil's predictions are highly speculative and often criticized. If any possibility exists that AGI will lead to a non-friendly Singularity, humans should proceed with caution. In a paper titled "Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood?" by AGI researcher and developer, Ben Goertzel, he argues in favor of constructing an intelligent but limited system designed with the sole purpose of delaying or preventing the development of a Singularity until we know how to create it in a positive way (2012).

AGI is, perhaps, not far from reality. Ben Goertzel is currently constructing something called Novamente. It's an integrative approach to AGI, and thus it brings together aspects of many prior AI projects and paradigms, including symbolic, probabilistic, and evolutionary programming as well as reinforcement learning approaches (Goertzel & Pennachin, 2007). Novamente is a learning computer, starting off like a child, useless, uneducated, with little information or cognitive ability, but eventually, it will grow up and learn about the world it inhabits about itself and about how it thinks. Goertzel's hope is that Novamente will be a huge breakthrough in AGI (Goertzel & Pennachin, 2007). Something like Novamente, which will be capable of learning and modifying its own software, could potentially lead to human level intelligence.

Figure 1. The exact future of AGI is still obscured in mystery. However, it is not merely science fiction, nor is it an insane fantasy of mad scientists.

The exact future of AGI is still obscured in mystery. However, it is not merely science fiction, nor is it an insane fantasy of mad scientists. While most AI research is currently focused on narrow AI, there exists a small group of serious scientists concerned with the more difficult problem of creating AGI. This ambitious work is as deserving of attention and respect as the other great scientific endeavors of our time. The consequences of this work may drastically affect the future of humanity, thus the claims made by these experts are deserving of realistic skepticism and well-thought out objections whenever they are applicable. The result of creating AGI or putting AGI in the hands of dangerous people could be severe. However, the benefits of creating friendly human-level intelligence may lead to exponential advances in many different fields, a greater understanding of how minds work, and a better world for the masses.

References

Arkin, R.C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. Artificial general intelligence, 2008: Proceedings of the first AGI conference. Eds. Wang, P., Goertzel, B., & Franklin, S. Washington DC: ISO Press, 2008. 51-62. Print.

Big Think (2010). After the Singularity, We'll All Be Robots [Video file]. Retrieved from http://www.youtube.com/watch?v=JR57633ztYc

DefenseTech.org (2011). Navy's second stealthy X-47B drone flies. DefenseTech.org. Retrieved from http://defensetech.org/2011/11/28/second-x-47b-uav-flies/#more-15485

De Garis, H. (2008). The Artilect War: A bitter controversy concerning whether humanity should build godlike massively intelligent machines. Artificial general intelligence, 2008: Proceedings of the first AGI conference. Eds. Wang, P, Goertzel, B., & Franklin, S. Washington DC: ISO Press, 2008. 362-373. Print.

Georzel, B. (2012) Should humanity build a global AI nanny to delay the singularity until it's better understood? Journal of consciousness studies, 19, 96–111.

Goertzel, B., & Pennachin C. (2007). Artificial General Intelligence. Springer.

Kurzweil, R. (1999). The age of intelligent machines. Cambridge, Massachusetts: The MIT Press.

Kurzweil, R. (2005). The Singularity Is Near: When humans transcend biology. New York: Penguin Books.

(Singularity Institute for Artificial Intelligence [SIAI], 2001, para. 2). 1.1: Seed AI. General Intelligence and Seed AI. Retrieved from http://singinst.org/ourresearch/publications/GISAI/paradigms/seedAI.html#glossary_crystalline

Wang, P., Goertzel, B., & Franklin, S. (2008). Artificial general intelligence, 2008: Proceedings of the first AGI conference. Washington DC: IOS Press.

Images Sources

http://zone14.files.wordpress.com/2010/04/080416-robot-doctor-022.jpg

http://www.imgbase.info/images/safe-wallpapers/tv_movies/the_matrix/14787_the_matrix_matrix_code.jpg

http://www.dayoftherobot.com/wordpress/wp-content/uploads/2011/07/DSC_3429.jpg