BIGMUN 2018
UNESCO – United Nations Educational, Scientific & Cultural

Research Report

Topic 2: The question of discussing the inhibitions, if any, which should be placed on the creation of self-altering algorithms

Emily Widjaja and Omar Majed

Introduction

The emergence of a malicious machine that will take over the world often sounds more like the plot line of a sci-fi film but with the development of self-altering algorithms, it is rapidly becoming a more realistic outcome. Self-altering algorithms, also known as self-modifying code, are algorithms that can modify itself and the lack of regulation and inhibitions surrounding these algorithms have become a cause of concern as evidenced in August 2017, when some 116 technological specialists sent an open letter to the United Nations, calling on them to "ban the development and use of killer robots."[1]

Currently, self-altering algorithms are more commonly used for debugging software and powering Netflix's recommendation engine than world domination. This research report will detail the issues surrounding the creation of self-altering algorithms.

Key Terms

Algorithm – A defined procedure or method that can solve a problem or perform a specific task. Specifically in Computer Programming, an algorithm is a code that defines the procedure and steps to solve a problem or complete a task.

Debugging – The process of identifying and removing errors from software.

Self-Altering Algorithms – A code that can change its own coding during execution of the code.

Terminal Value – the end goal that a software will continue working towards. When the terminal value has been reached, the algorithm is complete and will stop its execution.

Background Information

The rapid technological advancement has seen the development of technology with increasingly human capabilities. In previous years, one major distinction between human cognitive abilities and that of a machine's was the ability to learn. By taking mistakes and adjusting our behaviours to prevent them in the future, humans are able to learn. The introduction of self-altering algorithms, where a code can alter its procedures will allow learning to take place in a machine, but the quantity and contents of what it can or will learn are determined during the creation process.

For example, if the code has a terminal value of efficiency, let's say a specific value of time or steps in which the code must finish its execution, the code would alter itself and 'learn' with the ultimate goal of removing superfluous steps or procedure. This specification suitably lends its usage to debugging software and not destroying human life. However, if the code has a terminal value of world peace, this may or may not have a positive effect on human life. On one hand, it could learn that world peace

---

[1] Gibbs, S. *Elon Musk leads 116 experts calling for outright ban of killer robots.* The Guardian. Published: 20/08/17. Accessed: 06/01/2018. Available at: https://www.theguardian.com/technology/2017/aug/20/elon-musk-killer-robots-experts-outright-ban-lethal-autonomous-weapons-war

could be achieved by satisfying humanity's needs and wants, hence it would proceed to attempt to achieve world piece by improving production to increase the variety of goods available, for example. Alternatively, it could learn that in the absence of humanity or even all animals, there is no conflict and hence achieve world peace by wiping out all life except for plants.

The uses of self-altering code

Well-known examples of self-altering algorithms would be in Siri or the Netflix recommendation engine, used generally to make user experience more satisfactory. In medicine, they are utilised to improve the accuracy of a diagnosis by using algorithms to search databases to create a suitable treatment plan, modifying its code continuously to improve its abilities.[2] Currently, machines utilising self-altering code are still being used in conjunction with humans (for example, by a human doctor taking the information produced by the machine on the suggested treatments and then selecting the best one for the patient) but if it were to become fully autonomous, self-altering code and artificial intelligence can fill the shortage of medical expertise around the clock.

The extent to which self-altering algorithms are being used for militaristic purposes is unclear due to the lack of transparency surrounding military weapons. However, several countries, including China, Israel, South Korea, Russia, the United Kingdom and the United States, are developing and deploying partially autonomous armed drones, which may potentially utilise self-altering algorithms. Self-altering algorithms will most likely be necessary to make these weapons fully automated, as they would allow the weapons to learn and improve, and it is uncertain whether they can distinguish between a military objective and untargeted civilians.[3]

On a side note, self-altering algorithms are not necessarily synonymous with artificial intelligence even though artificial intelligence more often than not utilises self-altering algorithms. Artificial intelligence refers to software that can emulate human reasoning abilities, which can be accomplished with a normal code that is not self-altering, and does not necessitate that it has human cognitive abilities which would allow it to learn.

Major Countries and Organisations Involved

Campaign to Stop Killer Robots – An international coalition formed of multiple Non-Governmental Organisations (NGOs) working to ban fully autonomous weapons. While self-altering algorithms are not explicitly referred to, fully autonomous weapons would most likely utilise these algorithms.[4]

---

[2] The Medical Futurist. *Can An Algorithm Diagnose Better Than A Doctor?*. The Medical Futurist. Published: 2017. Accessed: 09/01/18. Available at: http://medicalfuturist.com/can-an-algorithm-diagnose-better-than-a-doctor/

[3] Human Rights Watch. *Killer Bots*. Human Rights Watch. Published: 2018. Accessed: 09/01/18. Available at: https://www.hrw.org/topic/arms/killer-robots

[4] Campaign to Stop Killer Robots. *About Us.* Campaign to Stop Killer Robots. Published: 2017. Accessed: 10/01/18. Available at: https://www.stopkillerrobots.org/about-us/

Facebook, Netflix, etc. – Self-modifying algorithms are used extensively by social media platforms and entertainment providers. Any inhibitions would be likely to affect these companies.

Relevant UN Resolutions

There are no relevant UN Resolutions on this topic.

Previous Attempts to Solve the Issue

Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be deemed to be Excessively Injurious or to have Indiscriminate Effects (with Protocols I, II and III) – This convention does not directly address any inhibitions involved in the creation of self-altering algorithms but under this convention, the use of a self-altering algorithm with a malicious terminal value would be banned for uses which had potential to harm the civilian population. The convention does not directly ban all usage; it only bans its use where the accomplishment of a military objective may result in an indiscriminate harm of the civilian population. Currently, the ability of self-altering algorithms to discriminate between unfamiliar humans is unclear, hence creating a loophole.

Possible Solutions

There are a few solutions available. Firstly, there is the potential for a complete absence of inhibitions or restrictions. On the other end of the spectrum, there is the possibility of a complete ban of all creation of self-modifying code. Neither option is particularly realistic due to safety and existential concerns in the case of the former and the setbacks in the case of the latter. In the complete absence of self-modifying code, technological advancement would slow immediately since they cannot be used to debug software, for example.

If one were to consider some inhibitions, a potential solution would be to allow the creation of self-altering code, but only for certain purposes, which would then create the question of what inhibitions and in what circumstances. For example, it may be preferable to have all self-altering code that could cause any physical harm banned and to allow all uses of self-altering code for medicinal purposes. However, some countries may seek to use self-altering code for military purposes, in which case, those countries would not endorse strict inhibitions relating to military uses of the algorithms.

If there were to be inhibitions, there is also the question of how to enforce these limitations. One such solution could be a global code approval panel composed of international experts but this solution would require that all member states be transparent on their uses of self-altering code. Alternatively, member states could monitor self-modifying algorithm creation domestically by working in conjunction with national ministries relating to information and technology and potentially having a United Nations team to audit. However, this solution may be difficult to implement for countries with less resources.

Bibliography

Campaign to Stop Killer Robots. *About Us.* Campaign to Stop Killer Robots. Published: 2017. Accessed: 10/01/18. Available at: https://www.stopkillerrobots.org/about-us/

Dvorsky, G. *How Artificial Superintelligence Will Give Birth To Itself.* io9. Published: 23/07/14. Accessed: 10/01/18. Available: https://io9.gizmodo.com/how-artificial-superintelligence-will-give-birth-to-its-1609547174

Dvorsky, G. *Can we build an artificial superintelligence that won't kill us?* io9. Published: 15/01/14. Accessed: 10/01/18. Available: https://io9.gizmodo.com/can-we-build-an-artificial-superintelligence-that-wont-1501869007

Gibbs, S. Elon Musk leads 116 experts calling for outright ban of killer robots. The Guardian. Published: 20/08/17. Accessed: 06/01/2018. Available at: https://www.theguardian.com/technology/2017/aug/20/elon-musk-killer-robots-experts-outright-ban-lethal-autonomous-weapons-war

Human Rights Watch. *Killer Bots*. Human Rights Watch. Published: 2018. Accessed: 09/01/18. Available at: https://www.hrw.org/topic/arms/killer-robots

The Medical Futurist. *Can An Algorithm Diagnose Better Than A Doctor?*. The Medical Futurist. Published: 2017. Accessed: 09/01/18. Available at: http://medicalfuturist.com/can-an-algorithm-diagnose-better-than-a-doctor/

Wheeler, T. *How social media algorithms are altering our democracy.* Medium. Published: 02/11/17. Accessed: 04/01/18. Available: https://medium.com/@Brookings/how-social-media-algorithms-are-altering-our-democracy-97aca587ec85