AI Weapon Systems and Our Future

Abdullah Barhoum
Normative Reasoning and Machine Ethics WS19/20
Student ID: 5041774
abdoo2@zedat.fu-berlin.de

December 14, 2019

Contents

0 Preamble

This following work is inspired heavily by the work of Amir Hussain[1] including his book *The Sentient Machine* [8] and his article *On Hyper-war* [7], and by Zachary S. Davis[2] in his article *Artificial Intelligence on the Battlefield* [5].

1 Introduction

Artificial intelligence (AI) has the potential to make many positive contributions to society, but it is important to avoid the negative effects of the use of AI. The use of AI by militaries in itself is not necessarily problematic, for example when used for autonomous take-off and landing, navigation or refueling. However, the use of AI to allow weapon systems to autonomously select and attack targets is highly controversial. In this paper we are going to discuss some of the concepts and misconceptions surrounding AI weapon systems, more specifically, the dangers it has on us as humans.

2 AI and Weapon Systems

One of the leading areas in which AI is being implemented at full force is in the military, and there have been some major advances, with two different perceptions regarding this subject.

At the one extreme, we have the idea that these AI systems, inspired by the fiction of Hollywood, might just become sentient and kill us all, which is how the general public fears AI. To any researcher in the field of AI, this is well beyond of what is feasible with the available technology nowadays.

On the other hand we have the national interests of some countries that would push the further development of such weapon systems to gain military advantage over competitors. The biggest concern remains the consequences of not only the use but even the development of such technologies. What if it creates a massive disparity?

The study of military history shows that generally stability is achieved when there is some sort of equality between leading nations. Having a competing power becoming dominant through these AI weapon systems creates reasons and gives impetus to instability, confrontation, and even wars.

This quickly escalates into the realm of game theory, where the assumption is, since it is not feasible to determine if the competitor is investing in AI weapons, it would be most beneficial to assume that they are and thus also invest to avoid reaching disparity. To make matters worse, certain countries are approaching AI with vim, vigor, and speed and making large investments, but certain other countries are not.

[1] Amir Hussain is the Founder & CEO of SparkCognition, and is on the Board of Advisors in UT Austin Computer Science

[2] Dr. Zachary S. Davis is a Senior Researcher at the Center for Global Security Research at Lawrence Livermore National Laboratory and a Research Professor at the Naval Postgraduate School in Monterey, California.

Nevertheless, one should not assume that AI technology itself, or its further development, is the threat. The human race is sufficient for that as it has been shown throughout history. Sophisticated technology in the hands of a highly malevolent individual is scary. The concern should not be AI, but when, how and why it is used.

2.1 Can't We Apply The Laws Of Nuclear Weapons?

During the nuclear age and cold war, the idea of "mutually assured destruction" dominated the decision space. More concretely, deciding to use any sort of nuclear weapon against other nations would only cause a response of the same damage, putting an end to the human race. This is not the case with AI.

Tackling this question from another point of view, the purpose of AI weapon system and the purpose of a nuclear weapon system are different. The only case in which nuclear weapons have been used was to break the will of a nation. On the other hand, the use of AI weapons is focused mainly on tactical victories or gaining certain advantages against the opponent.

In general, the trend is towards reducing the amount of kinetic firepower and increasing the amount of precision. Going back to any documentary about World War II would show huge numbers of bombers flying in and dropping large numbers of bombs to take down a bridge, or to take down a single building, losing a lot of people in the progress and creating mass damage in the area. The reason being the lack of accurate gun sights. But with AI it is possible to target very specific people within an armored division which is enough to collapse the fighting will of the soldiers.

Further distinction points between AI weapon systems and nuclear weapons include costs and verifiability, as the cost of manufacturing nuclear weapons is high, and creating such weapons would require building huge physical infrastructure to encapsulate and support the process, which could by detected by surveillance aircrafts or satellites. In a sense, the notion of "trust but verify" is valid for nuclear weapons.

In comparison, AI weapon systems are fundamentally software and cannot be verified this way. It is unrealistic to try and guarantee that a certain sequence of bits (and other permutations thereof) exists nowhere in any computer system in an entire country.

While the widespread and the accessibility of AI technologies have been huge factors in the advancement of the human kind, it also means that individuals with malicious intent can easily and cheaply access and use AI for harmful intents.

Orchestrated attacks across multiple actions and with multiple systems lead by autonomous controls and automatic targeting systems is not in the realm of fiction anymore, the risks that emanate from that kind of technology are visible, definable, and should be addressed as soon as possible.

3 The Destabilizing Aspects of AI

Zachary S. Davis, in his article *Artificial Intelligence on the Battlefield: Implications for Deterrence and Surprise* [5] summarizes the possible drawbacks of heavy reliance on AI in the military as follows:

3.1 Competition will bring Uncertainty

No one country can gain all of the benefits of AI while denying them to potential adversaries. Competition to gain advantage will bring uncertainty about the future military balance. AI may be seen by others as eroding mutual strategic vulnerability, thereby increasing the risk of war.

3.2 Data is Fragile

Perhaps the biggest obstacle to increasing reliance on AI is the problem of data reliability. AI systems are vulnerable to flawed data inputs, which can cause unintended consequences, as AI can magnify the "garbage in, garbage out" problem [12].

Data comes from many places and is not always carefully collected or curated. Compounding the problems with the data itself leading to skewed results, as AI often reflects human bias [6], or creates new biases based on flawed learning from the data provided [11] [9].

Transferring the inherent problems of data reliability and interpretation to the battlefield raises critical questions about the safety and reliability that accompany desirable attributes of speed and lethality. Accidentally hitting the wrong targets, for example, could have strategic consequences.

3.3 Countering AI Applications can be Straightforward

One of the most well known attacks is the *One pixel attack for fooling deep neural networks* [13], where the prediction of an image-classifying neural network can be manipulated by simply changing one pixel in the input image, furthermore, this attack can be deployed without knowledge of the underlying model.

The fact that AI is easily deceived invites efforts to counter the sought-after military benefits. By corrupting data in calculated ways, it may be possible to cause catastrophic equipment failures, miscommunication, confusion, logistical nightmares, and devastating mistakes in AI-reliant systems.

The "black box" problem of not understanding how and why AI makes decisions also means that it would be hard to recognize if data had been compromised to produce inaccurate outcomes, such as hitting the wrong targets.

3.4 Faster is Not Always Better

Speedy decisionmaking and operational execution may not serve well the goals of effective crisis management, as that same speed could be a disadvantage if it accelerates the conflict from crisis to war.

The battlefield advantages of AI-driven autonomous systems could shrink the time available for diplomacy to avoid or manage crises, because these systems tend to not include real-time reporting and analysis of national and international efforts to avoid, control, contain, or end a conflict.

3.5 Large Scale Complexity

AI-supported weapons, platforms, and operating systems operate according to custom-built software and hardware that is specifically designed for each separate system and purpose. Thus, it is necessary to integrate across scores of sensors, radars, weapons, and communications systems operating in multiple geophysical domains.

If this were not challenging enough, those systems would be built and operated by different agencies, commands, and contractors, with different authorities, access, and procedures.

The credibility of such "system of systems" must be called into question. The risks of setting up and maintaining such architecture invites challenges that could be destabilizing.

3.6 AI Unpredictability

The close operation and integration of multiple AI systems can be expected to have unanticipated results, some of which could have strategic consequences. It is uncertain how separate AI-infused platforms might interact with one another, as various AI-guided systems operate in shared battlespace.

With the internal "black box" mechanisms that produce AI outcomes, AI-to-AI interactions are likely to produce unanticipated and unexplainable results.

Close enough is not good enough when it comes to war; simulations or test runs in controlled environments are fully synthetic examples of what could happen on the battlefield, but do not reflect it fully.

Decisions of war and peace cannot, and will not, be left to predictive analytics. Unreliable data, machine learning bias, and interpretation contexts result in wider margins of error, which may be acceptable for scientific, economic, or logistic research purposes, but do not satisfy the practical and ethical demands of national security decisionmaking.

Machine learning cannot reliably predict the outcomes of elections or international conflicts, at least within margins of error acceptable when involving questions of war and peace. Much like self-driving cars, in which AI can correctly assess most, but not all, situations, a 90% success rate in military applications could mislead decision makers and put soldiers and civilians unnecessarily at risk.

3.7 Human in the Loop

The Department of Defense issued by then-Deputy Secretary of Defense Ashton Carter the following guidance [1]

> Semi-autonomous weapon systems [...] may be used to apply lethal or non-lethal, kinetic or non-kinetic force. Semi-autonomous weapon

> systems [...] must be designed such that, in the event of degraded or lost communications, the system does **not** autonomously select and engage individual targets or specific target groups that have not been previously selected by an authorized human operator.

In this case, a human has to make the ultimate decision, but this only holds in the U.S. Furthermore, as AI advances, it would continue to put more pressure on the OODA[3] loop, escalating the pace of war.

General (Ret.) John R. Allen[4] in an interview[5] at GLOBSEC[6] stated regarding this concrete problem:

> With AI, the [*OODA*] loop will shrink more and more until it almost collapses [...], I think we are going to have an issue with the human in the loop for some period of time. [...] When you competitor have taken the human out of the loop [...], they will be able to move faster. [...]
>
> **And it might be that we are driven at some point to a level of capacity and specificity within the algorithm itself, and we geo-fence the capabilities of that system to deliver ordinance in a way that gives as best as we possibly can the presence of the human in the loop.**
>
> We are going to find ourselves in in a hyper war environment where we are going to be selecting and specifically training officers to be able to exist in that environment, [...] where they have a bias for decisionmaking that can occur at high speeds.

The U.S. is willing to opt out of the human in the loop if that would lead to strategic and tactical advantages.

However, the International Committee of the Red Cross[7] has made a statement [3] in the *2019 CCW Group of Governmental Experts on lethal autonomous weapon systems* regarding this topic:

> The core issue is ensuring meaningful/effective/sufficient/appropriate **human control** over decisions to select and attack targets, **independent** of the technical sophistication of the weapon system.
>
> There is a need for human control in three key aspects:
>
> 1) Human supervision, and ability to intervene and deactivate
>
> 2) Predictability and reliability
>
> 3) Operational constraints, more specifically, constraints on the tasks, targets, environment, time, and scope.

[3] Observe, Orient, Decide, and Act

[4] John R. Allen is the president of the Brookings Institution, a retired United States Marine Corps four-star general, and former commander of the NATO International Security Assistance Force and U.S. Forces.

[5] https://youtu.be/UfIJGakA9v0?t=1214

[6] GLOBSEC is a non-partisan, non-governmental organization which focuses its activities on Security and Defense, Energy, Future of Europe and the European Neighborhood https://www.globsec.org/

[7] https://www.icrc.org/

Whether or not there are humans in every part of the decisionmaking loop, that loop is getting crowded.

The interface between humans and machines also raises critical questions about decisionmaking authority and organizational hierarchies [10]. Within the military, questions of rank, service branch, and responsibility for lethal actions can be contentious.

With scores of AI-informed battlefield systems operating at breakneck speed, each connected to its own chain of command, coordination among the humans who are in the loop of battlefield operations spanning multiple domains, agencies, clearance levels, and organizational cultures will be challenging.

3.8 Public-Private Partnerships

AI is freely available to everyone, it is being developed and applied beyond the reach of governmental controls.

As with many other dual-use technologies, governments rely on the private sector for the fundamental research and development, software, hardware, and expertise necessary for military AI use.

Many countries will use the same experts, companies, and global supply chains to support their military AI research and development, creating potential competitive conflicts of interest and security vulnerabilities related to the sharing of intellectual property.

Project Maven [2], one of the first examples of this partnerships, began in late 2017 by Google. It uses machine-learning to automatically label images, buildings, and other objects captured by cameras on drones, helping Air Force analysts identify unique targets.

Following months of protests from its employees, Google announced last summer that it would not renew its contract with the military, after thousands of Google employees, including dozens of senior engineers, have signed a letter protesting the company's involvement in this project [4].

4 Conclusion

AI weapon systems are a complicated problem. One would expect with multi-hundred stakeholders worldwide that every avenue of attack and every vector of thought has been applied.

There are many opinions and national interests, and those have not yet coalesced to a level that would allow preparation or come to any sort of a tangible conclusion on what approach to take with regards to these weapons systems in the global sense.

The *Ban Lethal Autonomous Weapons*[8], with more than 4500 AI researchers and tens of thousands of others, is trying its best to call for a complete ban on lethal autonomous weapons, alongside *The Campaign to Stop Killer Robots*[9].

Yet again, this relates back to the human condition, where we always strive for advantages over our opponents, which is not something that can be reprogrammed or deprogrammed from humanity in a near-term- or over the near-term period.

[8] https://autonomousweapons.org/

[9] https://www.stopkillerrobots.org/act/

References

[1] Department of Defense DIRECTIVE. https://fas.org/irp/doddir/dod/d3000_09.pdf. [Online; accessed 07-12-2019].

[2] Google Hedges on Promise to End Controversial Involvement in Military Drone Contract. https://theintercept.com/2019/03/01/google-project-maven-contract/. [Online; accessed 07-12-2019].

[3] ICRC statement, CCW GGE “LAWS”. http://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2019/gge/statements/25March_ICRC.pdf. [Online; accessed 07-12-2019].

[4] The Business of War: Google Employees Protest Work for the Pentagon. https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html. [Online; accessed 07-12-2019].

[5] Zachary Davis. Artificial intelligence on the battlefield. *PRISM*, 8(2):114–131, 2019.

[6] Jesse Emspak. How a Machine Learns Prejudice. https://www.scientificamerican.com/article/how-a-machine-learns-prejudice, 2019. [Online; accessed 07-12-2019].

[7] Amir Husain General (Ret.) John Allen. On Hyper-War. https://fortunascorner.com/2017/07/10/on-hyper-war-by-gen-ret-john-allenusmc-amir-hussain/, 2017. [Online; accessed 29-08-2019].

[8] Amir Husain. *The sentient machine: The coming age of artificial intelligence.* Simon and Schuster, 2017.

[9] Will Knight. Forget Killer Robots, Bias Is the Real AI Danger. https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger, 2017. [Online; accessed 07-12-2019].

[10] Tom Galvin Michael Piellusch. Is the Chain of Command Still Meaningful? https://warroom.armywarcollege.edu/articles/chain-of-command/, 2018. [Online; accessed 07-12-2019].

[11] ProPublica. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminalsentencing, 2016. [Online; accessed 07-12-2019].

[12] Hillary Sanders and Joshua Saxe. Garbage in, garbage out: how purportedly great ml models can be screwed up by bad data. *Technical report*, 2017.

[13] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *CoRR*, abs/1710.08864, 2017.