David Villalobos

Dr. Heather Fester

HASS 100 | Colorado School of Mines

22 November 2022

An Autonomous Weapons System Framework to Save Humanity

Targeted Stakeholders: AI Engineers & System Developers

Literature Review: Identifying Gaps in Literature

Stasis Level: Arguments of Proposal

As warfare technology development accelerates and quickly begins to outpace our control, we as engineers must be responsible for the safety of our creations and future generations. We must take a step back, reconsider, and must implement an autonomous weapons system framework to ensure that humans remain under the control of life-threatening decisions in order to guarantee the success of our future generations for which we are responsible. This framework however isn’t limited to autonomous weapons and the concepts should be implemented anywhere artificial intelligence is concerned.

PROBLEM DEFINITION

Artificial intelligence systems have begun rapidly advancing and gaining lots of attention all around the world as the Ukrainian war continues. Recently, many concerns have been raised over new fully autonomous weapon systems, their lack of safety with minimal regulatory statutes, and their severe danger to humanity; if an artificial intelligence regulatory framework to supervise autonomous weapon developments isn’t established and strictly upheld in the next few years, humans may go extinct in a few generations. The rate of our current technological growth is concerning and we must approach developments with caution. Ray Kurzweil is a world-leading computer scientist who graduated from Massachusetts Institute of Technology, he holds 21 honorary doctorates and honors from 3 U.S. presidents. He was the principal inventor of the first flatbed scanner, the first optical character recognition system, the creator of speech recognition systems, and much more. Kurzweil explains that technological advancements follow the Law of Accelerating Returns which means that new technologies have exponential development rates [1]. This exponential growth sounds fantastic, however extremely dangerous because of our lack of research into these new developments and implementations. Kurzweil introduced the idea of singularity as a point in time when we humans multiply our effective intelligence a billionfold by merging with the new artificial intelligent systems we have created [1]. Singularity has been estimated to be accomplished by the year 2045 which isn’t very far at all. Considering the technological advancements humans have made even in the last decade and the exponential growth, this is more than feasible which is why we must approach fully autonomous weapons and artificial intelligent systems with extreme caution and fear. Our technologies are advancing far too fast before we can even analyze and notice any potential consequences. It hasn’t even been 20 years since the first smartphones were created and nearly 85% of all humans have one. 20 years isn’t enough for us to analyze the potential hazards of having a small microwave in our pocket or the psychological impacts. Similarly, the exponential development rate of new technologies doesn't allow us time to even consider the potential hazards artificial intelligence and autonomous weapons could have on human existence. As Carl

Sagan said it best, advanced societies are never able to reach other civilizations because their innovations result in their extinction as they reach Kurzweil's point of singularity. If we neglect this critical decision-making, we will reach singularity and extinction at the same time, reinforcing this framework could be the buffer between human extinction and a flourishing world with interstellar travels.

Yoshaua Bengio is a world-renowned computer scientist, he is one of three founders of deep learning, an artificial intelligence system. Deep learning combines large amounts of data with multi-layered artificial neural networks to design the most sophisticated artificial intelligence systems that behave exactly like human brains. Deep learning is currently the most advanced form of artificial intelligence in which a system is fully autonomous and able to learn and make decisions without direct instruction from a human. Bengio explains that so many companies are using artificial intelligent systems irresponsibly. It’s shielded from the public, he explains, “A lot of what is most concerning is not happening in broad daylight. It’s happening in military labs, in security organizations, in private companies providing services to governments or the police.” [2] Bengio goes on to express his moral and security concerns with the use and abuse of autonomous weapons. As engineers we are responsible for our creations and the safety of our future generations, we must take a step back and consider these technologies. Bengio explains that we need international regulations because self-regulation, similar to voluntary taxation, doesn't work [2]. Enforcing this framework provides us the time buffer we engineers need to take a step back and re-evaluate the risk-to-reward ratio of implementing fully autonomous weapons and AI systems into our advanced technological world.

LITERATURE REVIEW (IDENTIFY GAPS IN LITERATURE)

To begin the in-depth analysis, we must first review what AI developers think about the implementation of autonomous weapons in a real war environment. Gauri Mishra is a computer engineer working at Cyfuture, she explains that by using autonomous military drones, we are taking a step in a very dangerous direction. Mishra explains that previous technologies such as radars would assist humans in warfighting but autonomous drones and weapons are the ones fighting for us. She forces the idea that there must be human facilitation when life-threatening decisions are made. The extreme dangers are exemplified by introducing an AI chatbot created by Microsoft; this chatbot spent less than 24 hours on social media and immediately began behaving like a “Nazi loving racist” [3]. Mishra also raises concerns about a lack of globally accepted morals; as we saw with the Holocaust, these autonomous robots may adapt to Eric Katz’s normative values [4]. Robots could also fall victim to psychological manipulation such as doubling explained by Robert J. Lifton, similar to the Nazis which allowed them to perform such evil acts while omitting the guilt [5]. The untethered autonomy creates instability and unpredictability of current intelligent systems that make them unfit for making life-death decisions. Philosopher Issac Taylor at Stockholm University is focused on implementing global justice into new technologies. He expands on Mishra’s idea by explaining that although they’re not fit for making life-death decisions, we can still harness their intelligence by throwing a human in the loop [6]. A human in the loop closes the responsibility gap and incorporates the emotions of a human into the decision-making. Emotions are the driving power behind a war, with no emotions there’s a lack of perception and depth contributing to the significance of the battle which AI systems lack. Philosopher Alex Leveringhaus is the coordinator for the Special

Interest Group for Ethics and Artificial Intelligence at Oxford University. He also agrees, explaining that humans have a range of emotions at war, “We need to leave space in warfare for pity, compassion, empathy, and the ability to put one’s gun down. Otherwise, we truly risk losing humanity in warfare” [7] he states. Artificial systems are incapable of experiencing emotions and a fully autonomous system would result in lowering the intrinsic value of life; “there must be a human in the kill chain” [7] he explains. We’re not alone with this idea however, more than 4,500 AI and robotic researchers have declared that AI should never be capable of taking human life. Leveringhaus sums up the idea, “Robots shouldn’t make decisions about life and death because they have no appreciation of the value of life” [7]. In our final framework, we must ensure that a human is included in the decision operation chain to invoke some level of trust in the users. This would also allow us to expand AI weapon autonomy in the future once it’s initially been introduced.

Trust is a major factor that is limiting the augmentation of autonomous weapons. Since these systems are brand new, humans' tendency is to approach them with caution similar to the discovery of nuclear weapons. Global Security Expert Heather Roff is a member of the Foreign Policy Program, her research is focused on autonomous weapons with international security and human rights protections. She explains that our research is getting ahead of itself and the actual fundamentals are being rushed [8]. The implementation of autonomous weapons will require a lot of trust and confidence which we heavily lack. From our understanding, she defines trust as several shades of gray contrary to our previous binary belief [8]. A lack of trust in autonomous weapons will lead to poor adoption rates in militaries or worse, the misuse of these autonomous weapons. Roff explains the different levels of robotic autonomy and calls attention to the direct correlation to human trust [8]. More predictable systems equate to higher human trust and vice versa. Roff explains that we must lower the autonomy of weapons and autonomous systems to gain trust [8]. She supports this by explaining, “present military structure views weapons as tools, [...]. War is a human activity, and trust in these systems is a necessary human element.”[8]. Researcher Ian Shaw has contrasting beliefs explaining that, “This ban should then be codified in an international treaty to prevent their development and procurement” [9]. However, Ian’s reason for this response could be due in part to the lack of trust we have in the current systems. As Roff explained, less autonomy results in more trust and this shows that the current developed systems have so much autonomy that researchers like Ian aren’t able to trust in these technologies [8]. As a scientific and engineering community we must agree to limit the autonomy of current weapons while we take time to grasp the true potential of these technologies. Having this small time buffer will allow us to fully process and acknowledge the limitations these systems should have before they’re expeditiously commercialized. Current autonomous weapon systems shouldn’t be capable of making life-death decisions due to their shallow understanding of the value of life, and lack of emotional comprehension. These current systems must incorporate a human in the loop not only to close the responsibility gap but also to ensure the safety of civilians in a war environment. In the future, we will expand AI autonomy as more research is completed and trust is gained.

FRAMEWORK (ARGUMENTS OF PROPOSAL - SOLUTION)

The ideal solution in this dangerous scenario would be for everyone to drop the weapons and discuss regulations. In fact, Sidney Axinn, a philosopher from the University of South Florida proposes that autonomous weapons must be put in the same category as chemical weapons and all nations must renounce their use [11]. He supports this by explaining that there's no honor in the killing of an enemy through a computer. Autonomous weapons are seen as cowardly wars since no lives are risked with the use of autonomous weapons. Unfortunately, the AI Cold War is rapidly accelerating and if the United States wants to keep up with all the competing countries, there's no time to slow our development. The National Commission on Artificial Intelligence is urging the US to accelerate research to remain competitive with Russia, China, and about 30 other nations designing advanced systems. Garcia Denise suggests that "The relentless pursuit of militarization does not protect us, we must focus on more important and cooperative issues such as natural disasters" [10] and while almost everyone would agree, nobody is willing to hold autonomous weapon research and globally regulate. Thus, similar to nuclear weapons, all nations must come together and acknowledge the extreme danger of negligently developing and commercializing autonomous weapons systems to abide by this framework that ensures our blooming future generations.

To maximize the prosperity of our future generations, this framework ensures that no autonomous weapon systems or artificial intelligence are capable of making life-threatening decisions. This is mostly due to our lack of technological advancements and our incapability of properly programming a value of life into computer code. Computers aren't truly able to understand life so it shouldn't be under their power to be able to remove it. Similarly, computers aren't capable of emotional comprehension, this once again is an aspect that should most definitely be considered in life-threatening decision-making. Both the value of life and emotional comprehension are consequential and crucial when life-threatening decisions are made, otherwise, human life risks losing intrinsic value since it can be taken so easily by a non-living robot. The simple solution to both of these missing characteristics is the involvement of a human in the loop! Incorporating a human into the final decision-making would also accredit war actions to a specific individual, this removes any blurs between accountability in war crimes and can be traced to a living person who is truly capable of understanding life.

While many would argue that this framework is actually working against the development of new technologies and purposely slowing our pace, this is completely true. Slowing the development of new technologies is the only way we as engineers and scientists are able to calculate the damages our designs could have. Since our creations are engineered so quickly, there's no time in the development phase to completely evaluate the effects on humanity. If we don't take a step back before these dangerous autonomous systems are commercialized, we could be making a catastrophic decision that could result in our extinction.

CONCLUSION

This framework addresses the major issues with autonomous weapons systems and artificial intelligence. It's our job now to address the human ego issues and come together as humanity to put the arms down by first acknowledging the extreme power we have at our disposal. Autonomous weapons should be taken with utmost seriousness just as any human killer is brought to justice. The Law of Accelerating Returns blinds our perception of technological growth, so much so that we’re not acknowledging the extreme dangers of these weapons [1]. This framework is the first step in renouncing these dangerous systems that could wipe out humanity. We were too close with nuclear weapons to our extinction, let's modulate these systems before we near extinction once again and ensure future blooming generations.

REFERENCES

[1] "Kurzweil Tracking the acceleration of intelligence," *Kurzweil The Law of Accelerating Returns Comments*. [Online]. Available: https://www.kurzweilai.net/the-law-of-accelerating-returns [Accessed: 28-Oct-2022]

[2] D. Castelvecchi, "Ai Pioneer: 'the dangers of abuse are very real'," *Nature News*, 04-Apr-2019. [Online]. Available: https://www.nature.com/articles/d41586-019-00505-2 [Accessed: 02-Nov-2022]

[3] G. Mishra, "Should Robots Be Allowed To Take Life-Death Decisions?," *Cyfuture Blog*, Jul. 25, 2019. https://cyfuture.com/blog/should-robots-be-allowed-to-take-life-death-decisions/[Accessed: 28-Oct-2022]

[4] E. Katz, "The Nazi Engineers: Reflections on Technological Ethics in hell - science and engineering ethics," *SpringerLink*, 16-Sep-2010. [Online]. Available: https://link.springer.com/article/10.1007/s11948-010-9229-z [Accessed: 14-Nov-2022].

[5] E. Filmus, "Doubling: Perpetrators and moral reconciliation," *Academia.edu*, 31-May-2016. [Online]. Available:

https://www.academia.edu/25738896/Doubling_Perpetrators_and_Moral_Reconciliation [Accessed: 14-Nov-2022].

[6] I. Taylor, "Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military‐Industrial Complex," *Journal of Applied Philosophy*, Nov. 03, 2020. https://mines.primo.exlibrisgroup.com/view/action/uresolver.do?operation=resolveService&package_service_id=11555582590002341&institutionId=2341&customerId=2340&VE=true[Accessed: 28-Oct-2022]

[7] A. Leveringhaus, "What's So Bad About Killer Robots?," *Exlibrisgroup.com* , May 2022. https://mines.primo.exlibrisgroup.com/view/action/uresolver.do?operation=resolveService&package_service_id=11555583350002341&institutionId=2341&customerId=2340&VE=true[Accessed: 28-Oct-2022]

[8] R. Heather, "'Trust but Verify': The Difficulty of Trusting Autonomous Weapons Systems," *mines.primo.exlibrisgroup.com*, 2018. https://mines.primo.exlibrisgroup.com/permalink/01COLSCHL_INST/1h4aoh8/cdi_informaworld_taylorfrancis_310_1080_15027570_2018_1481907 [Accessed: 28-Oct-2022]

[9] I. Shaw, “The Future of Killer Robots: Are We Really Losing Humanity?,” *http://eprints.gla.ac.uk*, Dec. 13, 2012. http://eprints.gla.ac.uk/73216/1/73216.pdf[Accessed: 28-Oct-2022]

[10] G. Denise, “Stop the Emerging AI Cold War,” *www.proquest.com*, May 13, 2021. https://search.proquest.com/docview/2528248164?pq-origsite=primo [Accessed: 28-Oct-2022]

[11] A. Sidney, “The Morality of Autonomous Robots,” *Exlibrisgroup.com*, 2022. https://mines.primo.exlibrisgroup.com/view/action/uresolver.do?operation=resolveService&package_service_id=11555583330002341&institutionId=2341&customerId=2340&VE=true[Accessed: 28-Oct-2022]

PROCESS NOTE

Proposing a framework for autonomous weapons systems was quite a daunting task. To begin, there are many starting points, this paper could have focused on proposing a solution to one of the aspects such as moral or legal but such would result in an extremely long and in-depth paper. Instead, it was decided to tackle the problem from a responsible point of view rather than an ethical or legal one that has many different points of view. Proving a responsible point of view is one that the majority of engineers overlook and is an aspect that isn't easily disagreeable upon. Providing a legal solution would be extremely difficult because there are too many moving parts that this paper could simply not cover. Similarly, an ethical framework would also be very convoluted due to the very biased and religious implications. Almost all of the lectures for NHV played an important role in developing a solution; while minimal lecture content was referenced, the background knowledge from all these lectures led to this final framework that encompasses almost all points of view and a strong ethical backbone. The most frustrating but contributing aspect is the source synthesis. It's very important to first understand what the scientific community thinks before blatantly giving a solution that may not even address an issue. This was originally a problem when providing a framework, at first it was providing a solution to a problem that doesn't carry weight in the community. Instead, it's more important to spend the extra time synthesizing all points of view to provide a solution that can truly make change and bring something new to the table. This was a very eye-opening experience which revealed how much content there is out there to review and the realization that we can never please everyone so it's best to provide a solution that helps out the greatest number of individuals following John Rawl's fundamental philosophy.