NATIONAL SECURITY FELLOWS PROGRAM

Ethical Imperatives for Lethal Autonomous Weapons

Dillon R. Patterson

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
JUNE 2020

National Security Fellowship Program
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org/NSF

Design and layout by Andrew Facini

Printed in the United States of America

NATIONAL SECURITY FELLOWS PROGRAM

Ethical Imperatives for Lethal Autonomous Weapons

Dillon R. Patterson

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

PAPER
JUNE 2020

About the Author

Dillon R. Patterson is a National Security Fellow at the Harvard Kennedy School of Government. Dillon is a Lieutenant Colonel in the Air National Guard, and has over 3,300 hours of combat experience piloting remote aircraft including the the MQ-1 Predator, the MQ-9 Reaper, and the RQ-170 Sentinel. He has earned a Bachelors of Science degree in Electrical Engineering from Embry-Riddle Aeronautical University, and a Masers of Arts degree in Defense and Strategic Studies from the United States Naval War College.

Table of Contents

A MQ-1 Predator and a MQ-9 Reaper assigned to the 432nd Aircraft Maintenance Squadron remain ready for their next mission at Creech Air Force Base, Nevada, May 5, 2015.

USAF Photo / Staff Sgt. Vernon Young Jr.

Introduction

In September 2017, Russian President Vladimir Putin addressed a group of students regarding the role of advanced technology in the future of warfare. During his address, Putin stated, "when one party's drones are destroyed by drones of another, it will have no other choice but to surrender."[1] Putin's remarks highlight a change in the character of war. Robots, often enabled by artificial intelligence (AI), represent an increased portion of military forces. In a November 2019 report to the United States Congress, the National Security Commission on Artificial Intelligence (NSCAI) identified Russia and China as utilizing various forms AI to advance their national agendas.[2] AI-enabled autonomous systems are tools that nations will not overlook in the development of national security plans.

The fields of automation and artificial intelligence are broad, having applications in diplomatic, informational, military, and economic activities. Within this realm, lethal autonomous weapon systems (LAWS) are a new enabler for achieving political ends through the application of the military instrument of power. As the world is past the point of considering whether robots *should* be used in war, the goal of the discussion herein is to examine *how* autonomous systems can be used ethically. This article seeks explicitly to demonstrate that fielding and employment of lethal autonomous weapons systems can be done effectively and ethically by maximizing the advantages and minimizing the shortfalls of both technology and the human mind.

In support of this position, the discussion will begin by defining autonomous systems and artificial intelligence. Additionally, a brief technical explanation of contemporary AI is provided. Next, the limitations and abilities of human cognitive capacity are reviewed to enable a comparison between humans and machines. The discussion then turns to Just

1 "Putin: Leader in Artificial Intelligence Will Rule World," CNBC, September 24, 2017, https://www.cnbc.com/2017/09/04/putin-leader-in-artificial-intelligence-will-rule-world.html.

2 Eric Schmidt and Robert Work, "National Security Commission on Artificial Intelligence Interim Report," November 2019, 11, https://www.epic.org/foia/epic-v-ai-commission/AI-Commission-Interim-Report-Nov-2019.pdf.

War Theory as a lens to provide classic ethical warfare frames when forming imperatives for the ethical use of LAWS in warfare.

Autonomy and Artificial Intelligence

Rationalizing positions for the ethics of LAWS requires a shared understanding of the technical concept in question between technologists, military leaders, policymakers, and ethicists. It is essential to understand that autonomy and AI are separate technical matters. Many autonomous systems in development incorporate some form of AI within their architecture; therefore, AI will be treated as an integral component for this discussion.

In a 2019 document addressing autonomy in future combat systems, Dr. Greg Zacharias, US Air Force Chief Scientist, borrows from the Merriam-Webster dictionary in defining autonomy as "the quality or state of self-governing; the state of existing or acting separately from others."[3] An autonomous system requires internal decision-making capability in-place of a human mind enabling the machine to utilize its network of sensors, information processors, and action nodes to *detect, decide, act, and update* itself as it operates within the mission environment. The introduction of a mission goal by a human initiates the sequence of autonomous operation. When the goal is simple, the decision mechanism can be simple. When the goal is complex, or the environment is dynamic, the decision mechanism must be complex. Thus, many autonomous systems have artificial intelligence at their core.

Unfortunately, no absolute definition of artificial intelligence exists. Massachusetts Institute of Technology professor Max Tegmark simplifies the matter by first defining intelligence as "the ability to accomplish complex goals."[4] Applying Tegmark's notion to non-human machines then yields a simple definition of AI as *the ability of machines to accomplish complex goals*.

3 Greg Zacharias, *Autonomous Horizons: The Way Forward* (Maxwell Air Force Base, Alabama: Air University Press ; Curtis E. LeMay Center for Doctrine Development and Education, 2019), 12.

4 Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, First edition. (New York: Alfred A Knopf, 2017), 50.

The AI employed within modern LAWS accomplishes complex goals through architecture that typically fits into one of two categories: logical processing (LP), and machine learning (ML).[5] LP systems require experts in a particular field to develop a mathematical model that defines an environment, such as weather patterns, traffic flows, or financial transactions. Computer scientists and engineers then utilize this model to program a set of exact instructions for the machine to follow when acting within the modeled environment. Machine learning takes a different approach. Instead of following instructions to act within a model, ML techniques begin with large quantities of data from the mission environment.

The machine uses data to discover trends or patterns that a human expert may never identify. Engineers shape the machine learning process by sending the data through a *training algorithm* that enables the machine to discover mathematical functions that approximately define the environment. Although the approximated functions may not be exact, they are typically more accurate than an expert-derived model because the machine can sort through far more data than a human mind.[6] Upon completion of the learning process, a new algorithm is programmed into the machine, which *activates the function* learned from the data.[7] The activation algorithm tells the machine how to utilize what it has learned as it operates in the mission environment to solve the goal presented by the human.

The most advanced contemporary AI systems are formed by combining multiple ML units into stacked layers, commonly referred to as *deep learning networks*. Each layer in the network is trained to learn a specific aspect of the target environment, providing a critical piece of the overall complex estimation of the environment.[8] Deep learning networks are so internally complex that engineers who shape the learning and activation algorithms can never know precisely how their machines come to the output actions, thus earning the nickname "black boxes."[9]

---

5 Terrence J. Sejnowski, *The Deep Learning Revolution* (Cambridge: MIT Press, 2018), 3.

6 John D. Kelleher, *Deep Learning*, MIT Press Essential Knowledge Series (Cambridge, Massachusetts: The MIT Press, 2019), 4–9.2019

7 Kelleher, 14.

8 Kelleher, 65–79.

9 Patrick Tucker, "Pentagon to Adopt Detailed Principles for Using AI - Defense One," February 18, 2020, https://www.defenseone.com/technology/2020/02/pentagon-adopt-detailed-principles-using-ai/163185

The high-powered analytical ability of AI-based autonomous systems will increasingly enable combat machines to accurately and expeditiously detect, decide, and act in battle. However, the manner through which AI is engineered restricts these systems to either the strict set of instructions given in a logical processing architecture or to activation boundaries of an approximated model created within a machine learning system. Ultimately, autonomous systems are *limited to action within the domain built into their decision mechanism*. The cost of autonomous high-speed precision and accuracy is domain inflexibility, which is a strength of the human mind.

The Human Mind

Although the design of many artificially intelligent systems powering LAWS are inspired partly by the human brain, they remain vastly different. Cognitive scientist Stellan Ohlsson asserts that complex brain activity utilizes multiple areas and levels of the neural cortex.[10] The complex nature of the world in which humans live requires the mind to operate along cognitive functions like perception, memory, thought, action, and learning.[11]

Through the process of encoding past experiences into episodic information, humans can reason, plan, and project future possibilities through what Ohlsson calls *monotonic learning*.[12] Learning in this manner is analogous to machine learning, in that patterns and trends are revealed within historical data. However, it is with monotonic learning that the similarities between the human mind and contemporary artificial intelligence halt. Ohlsson's model of *non-monotonic learning* describes how the human mind can "…suppress their past experiences and override its imperative for action."[13] This learning process is a result of the mind's ability to have mental conversion, creative thought, and adaptive outcomes.[14] Through

10 Stellan Ohlsson, *Deep Learning: How the Mind Overrides Experience* (Cambridge, UK: Cambridge University Press, 2011), 34.

11 Ohlsson, 34.

12 Ohlsson, 21.

13 Ohlsson, 21.

14 Ohlsson, 23.

non-monotonic learning, humans can exude intelligence by achieving complex goals in *multiple domains*, unlike the best deep learning networks.

While Ohlsson's writing explains how the mind functions and learns within a single domain and across domains to change conceptual beliefs, Psychologist Daniel Kahneman's work serves to illuminate *how well* the mind functions. Kahneman asserts that the mind has both an automatic function which makes decisions and directs actions during routine and emergency scenarios,[15] and an analytical function that overrides the automatic response to follow rules, perform calculations, and make choices after comparing options.[16] According to Kahneman, the analytical operation of the brain is prone to numerous limitations and performance reducing conditions.

Limitations like *ego depletion*, wherein high emotional strain correlates to depletion in physical stamina, or *cognitive overload*, when a high demand on the mind uses up so much mental energy that accuracy of thought and resistance to temptation can go unnoticed or uncared for,[17] degrade the analytical portion of the mind to check against inaccuracy in the automatic response. When the mind is hungry, tired, overworked, and emotionally drained, intentional thinking and reason diminish into the automatic responses the mind has developed.

Along with the concepts Ohlsson and Kahneman offer regarding human cognitive ability and limits, US Army Lieutenant Colonel (retired) Dave Grossman offers seminal research when seeking to understand how the mind reacts to lethal action. Grossman considers the propensity of humans to kill each other, the methods utilized to prepare soldiers to kill, and the resultant consequences within the individual and society.[18] Grossman asserts that contemporary military training has shifted from merely learning combat skills, such as how to fire a rifle accurately, to simulating the

15 Daniel Kahneman, *Thinking, Fast and Slow*, 1st ed., Harvard Library E-Reader Collection (New York: Farrar, Straus and Giroux, 2011), 33–35.

16 Kahneman, 35–36.

17 Kahneman, 41–42.

18 Dave Grossman, *On Killing: The Psychological Cost of Learning to Kill in War and Society*, Rev. ed. (New York: Little, Brown and Co, 2009).

combat environment, leading to more lethal conduct in battle.[19] However, the cost of highly competent human operators in battle has been moral injury when returning home.[20] An unintended and undesired consequence of creating better killers has been the dramatic rise of post-traumatic stress disorder (PTSD), and other mental illnesses.

The holistic picture of the human mind reveals an organic machine that is fallible, prone to errors in perception and judgment, influenced by sleep patterns, nutrition, workload, and moral temptation. It is flexible, and able to shift from one belief structure to another, despite robust history and habit. Finally, it is vulnerable to moral injury from the trauma of killing.

Understanding how the mind performatively compares to autonomous systems in combat should guide leaders in the establishment of policy for the employment of LAWS. Performance means not only combat success, but success in a manner that is morally acceptable to both the victor and the defeated. Comparing the ethical value of LAWS to human warriors requires objective criteria from which to compare the two; thus, the discussion turns to Just War Theory for the incorporation of classic ethical frames of warfighting.

19 Dave Grossman, *On Killing: The Psychological Cost of Learning to Kill in War and Society*, Rev. ed. (New York: Little, Brown and Co, 2009), 179.

20 Grossman, 43.

Just War Ethics

A classical paradigm for considering warfighting ethics comes from Just War Theory. Within this construct, the use of lethal force by states is divided into distinct categories of just cause for going to war, *Jus ad Bellum*,[21] and just conduct during war, *Jus in Bello*.[22] Although there are significant questions to be raised regarding the contribution of autonomous weapons in the initiation of armed conflict, the discussion herein is limited to the ethical employment of LAWS during war.

Canadian ethicist Brian Orend identifies several rules for just conduct during war, including discrimination and non-combatant immunity, proportionality, and no reprisals.[23] A just fight requires one's forces to accurately discriminate between targets that are of a legitimate military nature, and those which are not. Fighting should occur in a manner that does not bring harm to non-combatants.[24] Additionally, a just fight seeks effects that achieve their military objective in a manner that is proportional to their goal, not creating excessive damage, destruction, or suffering.[25]

Violation of *Jus in Bello* principles can and do occur during war. The stress of battle can cause the best warriors to make mistakes or intentionally act against the rules established to guide a just fight. Whether the cause of a *Jus in Bello* violation is fear, emotional or mental overload, physical fatigue, or group pressure, LAWS may function to prevent immoral acts in war as they are not subject to the stimuli which alter human decision making.

21 Brian Orend, *The Morality of War*, Second edition. (Peterborough, Ontario: Broadview Press, 2013), 33.

22 Orend, 111.

23 Orend, 138.

24 Orend, 112–13.

25 Orend, 125–30.

Ethical Imperatives for LAWS

Acceptable development and application of LAWS requires ethical concepts to chart the path forward. Guiding concepts should seek to maximize the advantages and minimizing the shortfalls of both technology and the human mind. Thus, the following five imperatives are offered together as an aid to shape a more ethical means of combat:

Imperative 1: Machines Do Tasks, Humans Exert Will

LAWS are limited by design to execution of a goal given by a human operator. Regardless of whether a machine utilizes logic processing, deep learning networks, or a complex AI stack with blended decision mechanisms, the machine is ultimately confined to detecting, deciding, acting, and updating itself within the confines of its design. For the foreseeable future, AI will likely not reach a level of development where machines have intelligence comparable to humans, commonly referred to as Artificial General Intelligence (AGI).[26] Although modern hardware supports robust deep learning networks, they continue to be built upon the same transistor-based micro processing technology. Many leading intellectuals in the AI field such as Max Tegmark,[27] Kai-Fu Le,[28] and John Kelleher[29] express some skepticism as to whether AGI will ever be possible. Thus, baring a major technological breakthrough, humans will continue to hold a significant advantage over LAWS through the ability to think broadly.

Although LAWS may be able to act and react faster and more accurately than a human for a given task, the underlying AI technology cannot think. Machines may be able to sense, detect, and act, but remain limited to the scope of analysis and action built-in by engineers. On the contrary, non-monotonic thinking ability has allowed humans to thrive in complex

26 Tegmark, *Life 3.0*, 52.

27 Tegmark, 132.

28 Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Boston: Houghton Mifflin Harcourt, 2018), 13.

29 Kelleher, *Deep Learning*, 241.

environments that are non-linear and dynamic[30] the ability to think in this manner is required to define the will achieved through a series of optimized tasks.

Imperative 2: Accuracy, Precision, and Proportionality

Jus in Bello principles of discrimination and non-combatant immunity are a function of *accuracy*, using military force solely against military targets, and *precision*, striking the intended target and avoiding non-combatants. The means through which a valid military objective was achieved most accurately and precisely provides maximal ethical results. Thus, LAWS are objectively more ethical than a human operator for a *particular combat task* if they are demonstrably more accurate and precise. As the technology matures, the role of human operators will likely move from directing machine actions, to consenting for machine actions, and ultimately to goal determination with the machine performing the entire detect-decide-act sequence.

Proportionality, applying only the necessary amount of force against legitimate military targets,[31] is similar to accuracy and precision in that quality of outcome matters. Because LAWS do not have human operators, cognitive bias associated with stress, fear, prejudice, or other factors are removable. Additionally, a human operator may need to accept undesired collateral consequences to non-combatants in order to achieve the military objective while protecting one's force from counterattack. LAWS must be utilized in a manner that achieves the desired military objective by reducing mistakes caused by cognitive errancy and through the ability to take greater risk than people, reducing collateral effects, providing a more proportional option.

The objective measure for ethics is in the ability of an autonomous system to perform a task accurately, precisely, and proportionally as *compared to the same act involving a human*. A realistic comparison between the human and the machine is necessary, as articulated in the third imperative.

30 Ohlsson, *Deep Learning*, 6.

31 Orend, *The Morality of War*, 125.

Imperative 3: Acceptable Error

Intentional acceptance of error is required to transition to a condition wherein LAWS operate as a normative part of the combat forces. The US government currently seeks to employ autonomous and artificial intelligence applications in national security that minimize adverse outcomes associated with machine errors. Literature published by the NSCAI,[32] the Defense Innovation Board (DIB),[33] and the National Institute of Standards and Technology[34] (NIST) provides expert guidance for developing systems that are safe, reliable, and trustworthy.

It would be a misstep however, to require near-perfect outcomes from autonomous systems as policy. The expectation of perfection is understandable considering historical approaches used to incorporate information processing enhancements to machinery. Most modern vehicles and factories use logic-processing to control system output and can be near perfect in action. However, LAWS are expected not just to act, but first decide on the correct action. Human expectations of the machine create the conceptual change in error tolerance. Autonomous systems represent a departure from the use of tools following a human decision to act, to those which decide and act in themselves.

Application of restraints requiring unrealistic quality measures well beyond human ability can potentially limit a means of warfighting that is ethically improved. The metric for error acceptance should be realistic, with objective measures that *compare the proposed autonomous machine against known error rates for humans under similar conditions*. Machines do not have to be perfect to be ethical, but they must be better than humans.

32 Schmidt and Work, "National Security Commission on Artificial Intelligence Interim Report."

33 Defense Innovation Board, "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Defense Department," October 2019, https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

34 National Institute of Standards and Technology, "U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools," August 9, 2019, 52.

Imperative 4: Accountability

LAWS can provide a means for enhanced accountability on the battlefield. Data acquisition is generally robust in autonomous systems, as data used by sensors to detect and decide on action is typically recorded and available post-mission. A human warrior's memory of events and details in stressful situations such as battle pales in comparison to the accurate record available within autonomous systems. Through data, leaders can reconstruct events creating transparency and accountability.

Accountability is not only enhanced within one's command chain, but across multilateral partnerships as well. Autonomous systems can be designed with *Jus in Bello* principles at the core of their decision mechanism, limiting the ability for *intentional violations* of rules of engagement. Weapon development should meet internationally accepted standards of ethics, attenuating an individual warrior's ability to misuse a weapon for an immoral act. Ethically designed tools of defense create an opportunity to project *Jus in Bello* values when selling military equipment to foreign countries, as well as use within one's force.

Imperative 5: Minimization of Moral Injury

LAWS should be employed in a manner that reduces the human risk for moral injury without sacrificing other *Jus in Bello* principles. In their article, *Avengers in Wrath: Moral Agency and Trauma Prevention for Remote Warriors*, authors David Blair and Karen House build on Grossman's work to present the case that propensity for moral injury increases as the human warrior develops a more intimate knowledge of a potential target. In many cases, combat roles like snipers or remotely-piloted aircraft operators, who perform long-endurance observation of a target before commencing lethal action, create a connection between the warrior and the target that can lead to moral injury.[35]

35 David Blair and Karen House, "Avengers in Wrath: Moral Agency and Trauma Prevention for Remote Warriors," Lawfare, November 12, 2017, https://www.lawfareblog.com/avengers-wrath-moral-agency-and-trauma-prevention-remote-warriors.

Intentional development and employment of autonomous systems offer the potential to break the psychological links leading to moral injury. Utilizing LAWS *for appropriate tasks* can lead to more ethical outcomes within both the justly acting force, and the non-combatant population enduring the rigors of armed conflict.

Conclusions

Regardless of whether a military act of violence is conducted by a human, through a human directing an automated weapon, or a fully autonomous system with no human involvement, the most ethical means of conducting a just war is that which maximally adheres to the principles of *Jus in Bello*. If a means of fighting is available which eliminates flaws due to human cognitive limitations, it should be considered as a viable alternative and ethically sound. Lethal autonomous weapon systems present a promising alternative to ethical warfighting by eliminating errors inherent in human monotonic thinking. Utilizing the human abilities of creativity, flexibility, and adaptability through non-monotonic thought to exert will through machines in many tasks may enable a more ethical means for armed conflict.

National Security Fellows Program
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org/NSF

Printed in the United States of America