LETHAL AUTONOMOUS WEAPONS SYSTEMS: MORALITY, OBLIGATIONS, AND RESPONSIBILITY IN THE LAW OF WAR

ABSTRACT

Emerging technology is disruptive and can reshape the world, sometimes in a manner that calls into question the concept of human agency. Artificial intelligence, specifically its application in Lethal Autonomous Weapons Systems (LAWS), represents one such technology. Instead of acquiescing to a world shaped by these systems, humans must be able to dictate the terms of use for emerging technology. This paper presents a set of proposals to address the shortcomings of current international legal regimes and develop solutions that will integrate autonomous weapons into warfare in a capacity chosen by stakeholders, not shaped by technology.

TABLE OF CONTENTS

INSETS WITH CITATIONS

MALD Thesis. Submitted by Kelly Crawford to Professor Dannenbaum on April 5, 2023 in full fulfillment of the MALD Capstone Requirement.

---

[1] "Deep Learning vs. Machine Learning - Azure Machine Learning," January 20, 2023, https://learn.microsoft.com/en-us/azure/machine-learning/concept-deep-learning-vs-machine-learning.
[2] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Third (Prentice Hall, 2009).
[3] Adam Zewe, "These Neural Networks Know What They're Doing," *MIT News | Massachusetts Institute of Technology*, accessed November 13, 2021, https://news.mit.edu/2021/cause-effect-neural-networks-1014; "Hyperparameters in Machine Learning - Javatpoint," www.javatpoint.com, accessed May 3, 2022, https://www.javatpoint.com/hyperparameters-in-machine-learning.
[4] "DHP-D258-Classification_and_Regression_with_KNN.Ipynb," accessed May 3, 2022, https://colab.research.google.com/drive/1PTFpYeNFbyUY0Ig_gXrpADjRaEnXiL3A?usp=sharing.

I
INTRODUCTION

"A weapon system that, once activated, can select and engage targets without further intervention by a human operator."[5]

-United States Department of Defense (DoD)

"[A weapon system] capable of understanding higher-level intent and direction. From this understanding and its perception of its environment, such a system can take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be."[6]

-Stockholm International Peace Research Institute (SIPRI)

Lethal Autonomous Weapons Systems (LAWS) are problematic for the current legal regime surrounding warfare. No matter the position you take on their efficacy or the morality of using artificial intelligence (AI) during the targeting cycle, both International Humanitarian Law (IHL) and the International Criminal Justice (ICJ) system do not emphatically address the prospect of a non-human system taking part in decision-making. This paper will first address if LAWS have moral agency and if they can carry out IHL obligations. Next, it will address the shortcomings of the current IHL regime and propose a new framework to accommodate LAWS. Then, it will dissect the prospect of prosecuting a war crime "committed" by an autonomous system and propose a new mode of liability to account for LAWS. Finally, it will address the pragmatic concerns of introducing LAWS onto the battlefield.

LEGALITY AND RESPONSIBILITY

Discussing the regulation of LAWS requires acceptance of a set of assumptions. First and foremost, LAWS are legal under IHL. While the topic is not without controversy given the

[5] United States Department of Defense, "Autonomy in Weapon Systems," 3000.09 § (2012).
[6] Vincent Boulanin, "Mapping the Development of Autonomy in Weapons Systems" (Stockholm International Peace Research Institute, November 2017), 6.

increased attention paid to LAWS by the Convention on Certain Conventional Weapons (CCW), it seems that there is at least a concern, if not acceptance, that IHL does not outlaw autonomous weapons *per se.* Schmitt states, "whereas some conceivable autonomous weapon systems might be prohibited as a matter of law, the use of others will be unlawful only when employed in a manner that runs contrary to [IHL]."[7]

The second assumption is that enough consensus exists that there will not be a "black hole" of legal accountability. Detailed reports from the Harvard Law School Program on International Law and Armed Conflict (PILAC) and SIPRI outlines this issue.[8] The PILAC report states, "In sum, [International Criminal Law] and, especially, IHL already address many of the concerns raised in relation to [Autonomous Weapons Systems]—but ICL and IHL may not be sufficient to address all of those concerns."[9] The SIPRI report raises questions regarding these shortfalls,

Deep learning, machine learning, and AI

- **Deep learning** is a subset of machine learning that's based on artificial neural networks. The *learning process* is *deep* because the structure of artificial neural networks consists of multiple input, output, and hidden layers. Each layer contains units that transform the input data into information that the next layer can use for a certain predictive task. This structure allows a machine to learn through its own data processing.
- **Machine learning** is a subset of artificial intelligence that uses techniques (such as deep learning) that enable machines to use experience to improve at tasks. The *learning process* is based on the following steps: 1) Feed data into an algorithm. 2) Use this data to train a model. 3) Test and deploy the model. 4) Consume the deployed model to do an automated predictive task. (In other words, call and use the deployed model to receive the predictions returned by the model.)
- **AI** is a technique that enables computers to mimic human intelligence. It includes machine learning.

[7] Michael N. Schmitt, "Autonomous Weapons Systems and International Humanitarian Law: A Reply to the Critics," *Harvard Law School National Security Journal* Journal 4 (2013): 1–37.
[8] Dustin A. Lewis, Gabriella Blum, and Naz K. Modirzadeh, "War-Algorithm Accountability" (Harvard Law School Program on International Law and Armed Conflict, August 2016); Vincent Boulanin, Netta Goussac, and Laura Bruun, "Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human–Machine Interaction" (SIPRI, June 2021).
[9] Lewis, Blum, and Modirzadeh, "War-Algorithm Accountability," 101.

including concerns over when and how much legal responsibility should be attributed to a person or state.[10] For example, should an engineer who designed the system be held partially responsible for an IHL violation, or does the commander who employed the system have (or should have) the appropriate knowledge to be individually accountable?

While the current regime surrounding state and individual responsibility is glaringly imperfect, there is no reason to think that it will collapse with the introduction of LAWS into battle. Properly addressing the legal shortcomings requires answering the question, "Can artificially intelligent agents be permitted or required to perform IHL obligations?"[11]

[10] Boulanin, Goussac, and Bruun, "Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human–Machine Interaction."
[11] Boulanin, Goussac, and Bruun, 55.

II
MORALITY AND IHL OBLIGATIONS

MORAL AGENCY

Moral concerns permeate every conversation about LAWS. Through 2020, around 30 countries and 165 non-governmental organizations cited ethical concerns in calling for outright bans on LAWS use.[12] Many states voiced concerns at the 2015 and 2016 CCW Meeting of Experts on Laws. Pakistan stated that "LAWS create an accountability and transparency vacuum…If the nature of a weapon renders responsibility for its consequences impossible, its use should be considered unethical and unlawful."[13] Argentina went further, stating, "at the current moment in its evolution, [IHL] still does not give solid answers to the challenges laid out by an autonomous system that will become able to make the decision to take a life, with complete independence from a human order."[14] Costa Rica called for humans to be the sole deciders of the use of lethal force. At the same time, Australia questioned "whether the principles of humanity and dictates of public conscience can ever allow machines to select, attack and kill human beings, entirely outside human control."[15] Ireland may have encapsulated the humanity concern the best: "the decisive questions may well be whether such weapons are acceptable under the principles of humanity, and if so, what conditions."[16]

The "decisive question" is one of moral agency. Do machines possess it? In a piece discussing if artificial intelligence could make ethical decisions and what that would look like, Ross Bellaby writes, "for an AI weapon to make a truly independent decision would require it to

[12] Kelley M Sayler, "Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems" (Congressional Research Service, n.d.), 2, https://crsreports.congress.gov.
[13] Lewis, Blum, and Modirzadeh, "War-Algorithm Accountability," 138.
[14] Lewis, Blum, and Modirzadeh, 138.
[15] Lewis, Blum, and Modirzadeh, 147.
[16] Lewis, Blum, and Modirzadeh, 146.

have autonomy, and with it the ability to reflect upon all the input and parameters it is given, with the potential to disagree."[17]

For a machine make or "mimic" moral decisions, it would either need to have pre-programmed parameters known as a "top-down" approach or go through a machine learning process, or a "bottom-up" approach. Neither one of these meets Bellaby's requirement for reflection. No matter how advanced a machine may seem, it still does not possess moral agency.[18] Bellaby's argument around moral agency does not rule out the technical possibility of IHL obligations being pre-programmed into a machine, but instead emphasizes they will be algorithmic decisions rather than moral ones.

Moral agency implies a cost or burden associated with a decision, and programming moral decisions into LAWS would remove this burden, creating an ethical problem. Three parties bear the burden of war: the decision-maker, the attacker, and the target.[19] But when LAWS make decisions and take actions for the first two parties, automating their role in war, it also relieves them of their moral burden. With no moral burden to bear, the cost of war falls solely on the target.

Removing moral authority from a combatant (since a machine has none) will significantly cheapen the cost of war. This is not financial cheapening but moral cheapening. If there is no moral cost levied on society to engage in hostilities, why should society avoid them? Asymmetry in combat has increased since the Vietnam War, and the technological capability to conduct near-casualty-free operations were driving factors in strategy in the 1991 Gulf War and NATO missions in Kosovo. With no moral burden associated with war and the risk of casualties near zero, the only

[17] Ross W. Bellaby, "Can AI Weapons Make Ethical Decisions?," *Criminal Justice Ethics* 40, no. 2 (May 4, 2021): 87, https://doi.org/10.1080/0731129X.2021.1951459.
[18] Bellaby, "Can AI Weapons Make Ethical Decisions?"
[19] The separation of the "decision-maker" and "attacker" will be circumstantial. A scenario where they are the same person holding the rank of an E-1 is equally as plausible as one having multiple decision makers comprising the entire chain of command up to the Commander in Chief. I will use the term "combatant" to account for both these individuals for a) clarity and b) the potential combination of the two jobs within circumstances using LAWS.

firing break left is the price tag. Outside of the monetary value of the LAWS, there is little society must contend with in weighing military action.

If the use of LAWS causes the problematic removal of a combatant's moral burden, then what should we expect from a combatant? What about from military institutions? The argument against relieving a combatant of their moral obligation by way of LAWS is not meant to usher in a universal warrior ethos. Nor is this argument meant to establish a threshold of danger for a combatant. It would be unfair to ask a combatant or a military to forego technology or tactics simply based on asymmetry. If a pilot thinks it is unfeasible to fly below a certain altitude for target verification because of enemy air defense systems, then that is a choice the pilot or their commander has the agency to make. It may or may not be legal or ethical, but the burden they will shoulder for that decision is moral, and as moral agents, is their own.[20] These decisions, reflected as obligations in IHL, are where the burden of war lies. The international law community needs to clarify what IHL obligations are appropriate for a machine to accomplish and which, if any, should be reserved for a human.

THE DECISION CYCLE

If LAWS lack moral agency, do they also lack the ability to carry out International Humanitarian Law obligations? Under IHL, certain obligations must be met for a killing to be lawful. These obligations are cumulative and are:

1) Distinction between combatants and non-combatants[21]

[20] An emerging argument cited in *Beyond the Ban: Comparing the Ability of "Killer Robots" and Human Soldiers to Comply with IHL* by Lena Trabucco and Kevin Jon Heller, and *Robophobia* by Andrew Keane Woods, as well as many other articles, raise the prospect that we may be asking too much from robots—holding them to a higher standard than a human. This leads to questioning if the moral thing to do is to *use* the technology as a way of lessening death. This argument may be a *factor* in the moral argument, but it does not address the issue I raise in this paper surrounding a moral burden.

[21] "Protocol Additions to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (AP/I)" (1977), art. 50.

2) Protection of civilians from indiscriminate attacks[22]

3) The necessity to take feasible precautions to avoid civilian losses[23]

4) The proportionality of an attack[24]

Additionally, a fifth action pertinent to the discussion of moral agency should be added to this list:

5) The act of killing, or "batteries release."[25]

Each of these obligations must be met in totality for an attack to be legal and they are inseparable from each other. However, in evaluating what may be possible for LAWS to accomplish under IHL, each obligation needs to be evaluated independently. This evaluation should consist of a two-part test. First, is it technically possible for LAWS to carry out this obligation. Second, is there an inherent moral burden associated with the obligation and would thus require a human with moral agency to carry out the obligation.

Before proceeding into analyzing each obligation, the issue of doubt needs to be addressed. Doubt, or uncertainty, can attach to any obligation. A combatant can doubt, for example, if a person is a civilian or not. Under IHL, if this doubt exists then "that person shall be considered a civilian."[26] In this scenario, an autonomous weapon's uncertainty would be characterized mathematically as a quantifiable error or accuracy rate i.e., the LAWS determines an individual is a combatant with an error rate of 4%. If LAWS is less than 100% accurate that a person is a civilian, then uncertainty, and possibly doubt, exists for that LAWS.

---

[22] Protocol Additions to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (AP/I), art. 51.
[23] Protocol Additions to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (AP/I), art. 57.
[24] Protocol Additions to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (AP/I), art. 51.
[25] "Batteries release" is the decision to kill at the final most point in the kill chain. Among other things, this can manifest in the pulling of a trigger, pushing a button, or entering a line of code. It has no geographic limitations.
[26] Protocol Additions to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (AP/I), art. 50.

This is a marked difference from conventional definitions of doubt in two ways. First, what distinguishes the nature of accuracy in LAWS as compared to the nature of accuracy in a conventional weapon like a howitzer, is that accuracy in LAWS is used in a decision-making capacity. A howitzer does not "distinguish" civilians, take "feasible precautions," or partake in any other IHL obligation. These are precisely what AI is meant to automate.

What distinguishes uncertainty in LAWS in comparison with a human is that we can measure its uncertainty mathematically, something we cannot do with a human combatant. Does this mean that uncertainty should equate to doubt? If so, that means any LAWS with less than 100% will be in a constant state of "doubt" and needs to be programmed accordingly to act in accordance with international law. The other option is that a threshold is established establishing an appropriate level accuracy, and if that accuracy is not reached, then the LAWS is legally "doubtful." It could further be argued that LAWS deployed without sufficient accuracy may be illegal *per se*.[27]

The Problem of Scaling Logic & Fail Cases

"The world is composed of facts that do or do not hold in any particular case."

- When interpreting partial sensor information, a logical agent must consider *every logically possible* explanation for the observations, no matter how unlikely. This leads to impossibly large and complex belief-state representations.
- A correct contingent plan that handles every eventuality can grow arbitrarily large and must consider arbitrarily unlikely contingencies.
- Sometimes there is no plan that is guaranteed to achieve the goal—yet the agent must act. When it acts, what happens? These, "fail cases" could result in a multitude of options with legal and moral tradeoffs.
- Examples of "fail cases" could be to notify a human, system crash, random act, or hallucinations

[27] A question arises that if LAWS is less than 100% accurate and has doubt, then a moral dilemma will invariably arise and then moot the entire discussion of technical feasibility. This moral dilemma, however, is not always *inherent* in every obligation—the criteria for the second test. If LAWS are deployed with inferior technical capabilities, then they would be illegal to use in the first place.

Further discussion of the topic of accuracy "thresholds" is in Section IV. In the following analysis of IHL obligations, the concept of "doubt" *vis-à-vis* LAWS will be considered as a technical question that will need to have been answered for the weapon to be deployed legally in the first place. If the IHL obligation can be executed through a technological process, whether that process is *via* machine learning, sensors, or something else, then that will constitute the passing of the technical test that obligation.

The first two obligations, "distinguishing between combatants and non-combatants" and "protecting civilians from indiscriminate attacks" are similar in that they are technically feasible to accomplish and do not have any inherent moral burden. If analyzed independent from the other obligations, these two obligations do not raise any inherent or unavoidable moral risks, if the technology used functions as designed.

Distinction requires only sensors and machine learning to interpret the sensor data. Like autonomous vehicles, LAWS need input data from radars or visual sensors to run a classification process. Instead of "pedestrian" or "vehicle," it would distinguish between "combatant" or "non-combatant."

There are obvious technical limitations, but no reason why anyone would need to be morally burdened by these decisions. The assumption that the technology can properly execute the distinction task within certain specifications is crucial. If it cannot, the faulty technology will prompt questions about the validity of its deployment in the first place. However, capable technology will not require moral agency in what boils down to a task of classification. For example, if a LAWS has sensor input data of a person and that LAWS wrongly determines that person is a combatant, then that is a technical failure. This technical failure will rightly raise other questions about the legality of that particular system and moral questions about where the technical

threshold for accuracy should be, the obligation does not, however, have an inherent moral burden.

Likewise, "the protection of civilians from indiscriminate attacks" is an obligation that rests on technical capabilities. A particular type of munitions or weapon will or will not cause collateral damage. The target is either a military object or not. While differing scenarios will present differing levels of complexity, these are not necessarily "decisions" as much as they are rules and recognitions. In arguments for the use of AI in warfare, proponents find their arguments in precisely these decisions. A machine that can recognize non-combatants or recommend munitions tailored to a specific combat scenario has the potential to reduce collateral damage significantly. While there are other intertwined aspects of the decision cycle that have inherent moral burdens, distinction and protecting civilians from indiscriminate attack, if analyzed independently, do not. Because of this, these obligations could be accomplished by a machine, independent of a human's moral agency.

The third obligation, "the necessity to take feasible precautions to avoid civilian losses" is also technically feasible to be carried out by a machine, but because of the inescapable need for a human to be present at some level in the chain of command to make the initial decision to deploy LAWS, there is an inherent moral burden.

The feasible precautions taken by LAWS would be similar to the first two obligations. Sensor input data can be used to verify there are no non-combatants, and then an automated process can pick the means and methods of attack. As with first two obligations, this task has technological limitations and raises the same threshold questions. A major difference, however, is inescapable human involvement. It is impossible and far beyond the current reality to automate the entire command chain up to the Commander in Chief. At some point, a human must make a decision and order the deployment of the automated weapons system. This decision may rest at the tactical level

or at the strategic level, and depending on the level the questions arising will look very different. However, contained in each decision is a moral one that involves weighing the information at hand against risks to life and what value that life has in the context of war, be it a tactical scenario or a strategic one. Although this obligation may be possible to be accomplished by LAWS, depending on the circumstance, the inherent moral burden associated with the deploying of these weapons require human decision making at some point in the decision cycle to accomplish this obligation in good faith.

The fourth obligation is "proportionality." The problem with proportionality is that there is an expectation of causing the loss of civilian lives or objects.[28] For LAWS, this decision would require assigning mathematical value to civilian lives. While assigning mathematical value to human lives is certainly possible, this is an inherently moral decision. This obligation does not rest on the technical ability of LAWS in the same way that distinction does. Even if an engineer were able to develop a system that could conform and adjust to changing political and military objectives during a military campaign and understand the value of a military target in the context of military and political objectives, there is no way to circumvent the weighing of lives against these objectives. From the outset, it is expected that the attack may kill non-combatants, and in killing there is a moral burden to bear, something LAWS cannot do, this obligation has an inherent moral decision and cannot be carried out by LAWS.

The final decision is "batteries release," the last firing block in the cumulative test: the decision to kill. This decision can take many forms, from the physical act of pulling a trigger to an input to an autonomous system. It is essential that this decision is not construed to come earlier in the cumulative test process. Presumably, the decision to kill was made earlier in the planning

[28] Protocol Additions to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (AP/I).

process or in the decision to deploy an autonomous weapon. The point of keeping this decision with a natural person at the final stage is that no matter how many inputs a machine receives or what quality they may be, the outcome of that decision will be mathematical, logical, and final.

Systems already in place may mimic the hyper-logic of a machine. There is already a very systematic, almost computational process of dubious legality involved in "signature strikes."[29] The difference lies in in choice and agency. A human can *choose* to weigh the final act of taking a life as a moral and subjective problem, not a mathematical one. A human has the moral agency and capacity to make these decisions. A human will ultimately bear the burden of them.

In summary, the obligations of distinction and the protection of civilians from indiscriminate attacks are the only obligations that have the potential to be both technically feasible for LAWS to carry out and have no inherent moral burden. Of course, it is possible to construct a scenario where these obligations do have a moral burden, but that will be contextually specific. The purpose of independent evaluation of IHL obligations is to determine what the technical and moral *potential* is for LAWS in combat, not to construct a universal hard fast rule for their use.

However, the problem with this analysis is that it is not realistic. These obligations in the decision cycle are cumulative and temporally compressed—the whole cycle can happen in as little time as it takes to pull a trigger. What it does do, however, is begin to establish limitations as to what machines shouldn't do autonomously—any obligation that has an inherent moral burden. Once these limitations are established, then we can begin to decide where and how human control can be established.

[29] Kevin Jon Heller, "'One Hell of a Killing Machine': Signature Strikes and International Law," *Journal of International Criminal Justice* 11, no. 1 (March 1, 2013): 89–119, https://doi.org/10.1093/jicj/mqs093.

THE (PARTIAL) FALLACY OF MEANINGFUL HUMAN CONTROL

The principle of meaningful human control inserts human reasoning into the decision-making of a machine. In this case, human control would come at points in the decision-making cycle that the LAWS is unable to accomplish, either morally or technically. Yet meaningful human control remains ill-defined and, further, may not even be possible.

Returning to the PILAC report, multiple states attempt to define meaningful human control, sometimes outright or sometimes in reference to other aspects of the accountability regime. Both the United Kingdom and Poland reference accountability in this way, with the U.K. stating, "…there must always be human oversight and control in the decision to deploy weapons…Responsibility will flow up the chain of command, which is so important in military structures."[30] The United States is more flexible in its approach by allowing commanders and operators to arrive at an "appropriate level of human judgment."[31] Furthermore, the United States notes that "'human judgment over the use of force' does not require manual human 'control' of the weapon system…but rather broader human involvement in decisions."[32]

The Netherlands brought a more nuanced approach in a statement to the Group of Governmental Experts on LAWS in 2019. It said that meaningful human control "…should be understood within the context of design, development and operational use…" and acknowledged that the process of targeting is complicated and involves human operators at various levels, and therefore the introduction of LAWS would not drastically alter the process. However, they did state that weapons that "can change their goal-function independently or alter pre-programmed

[30] Lewis, Blum, and Modirzadeh, "War-Algorithm Accountability," 139.
[31] Sayler, "Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems."
[32] Sayler.

conditions and parameters" are not under meaningful human control.[33]

Each of these definitions is distinct and none dives into the pragmatic realities of human control. There are powerful arguments that human control is not even a possibility. In challenging human control, Schwarz raises concern that constant interaction with autonomous technology may very well lead to a "moral deskilling," especially if viewed as a proper substitute for human agency.[34] The arguments made in this paper embrace an "instrumentalist position" of LAWS and are accepting of the idea that this technology is a "tool" and will therefore always be subordinate to a human.[35] The consequence of this position, according to Schwarz, is an underestimation or ignorance of how a LAWS could shape a combatant's decisions.[36]

"Moral deskilling," however brilliantly outlined in Schwarz's paper, is not a new phenomenon. Although the degree to which humans trust, and integrate themselves, in technology systems may amplify this effect, humans constantly must fight the inclination to "take the easy way out," regardless of the technological sophistication of the decision-making. Daniel Kahneman's book, *Thinking Fast and Slow*, published in 2013 and lacking discussion of the new issues surrounding AI, explains two modes of thinking. System 1, an automatic system of thinking, and System 2, a critical system of thinking.[37] It takes effort for people to switch to System 2 thinking and overcome biases, intuition, or perhaps, trust in a machine. None of this is meant to denigrate the danger of moral deskilling or the likelihood that human control will devolve into "rubber stamping" a machine's decisions, but only should show that social psychologists have

[33] de Jongh, "Statement of the Netherlands" (Statement, Group of Governmental Experts on LAWS, Geneva, April 26, 2019).
[34] Elke Schwarz, "Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control," *Philosophical Journal of Conflict and Violence* 5, no. 1 (May 20, 2021): 69, https://doi.org/10.22618/TP.PJCV.20215.1.139004.
[35] Schwarz, 55.
[36] Schwarz, 55.
[37] Daniel Kahneman, *Thinking Fast and Slow* (Farrar, Straus and Giroux, 2013).

been grappling with this issue for a long time.

In another book, *Sources of Power: How People Make Decisions,* Gary Klein analyzes what makes a good decision maker in high-stress, time-compressed situations.[38] Klein uses a study on firefighters, a job that is analogous to a soldier in terms of high-stakes, high-stress decisions, and walks through how they succeed, as opposed to fail.[39] The upshot of this research is that in situations where system 1 would seemingly dominate, say in the heat of battle, decision makers succeed in part by drawing on a wide array of non-analytical decision making tools such as "—the power of intuition, mental simulation, metaphor, and storytelling."[40] The question then may not be if LAWS will amount to the elimination of agency, but how can humans incorporate their current capacity to make tough, high-stress, high-stakes decisions into a new sophisticated tool without losing it.

The moral concerns of LAWS and Schwarz's critique, can be addressed by institutionalizing of the idea of moral agency and moral expectations of the combatant. Trying to shoehorn humans into every decision point in an autonomous weapons system is not the correct answer. In one extreme, combatants lose all agency themselves and "rubber-stamp" decisions made by LAWS. On the other end, the technology becomes burdensome and unworkable. The best solution at present is to codify the moral boundaries for LAWS laid out in this paper, establish an accountability regime that reinforces these moral obligations through command responsibility, and establish standards in the development of tech that allows for a better human understanding of LAWS.

[38] Gary Klein, *Sources of Power: How People Make Decisions* (MIT Press, 1999).
[39] Klein, chap. 1.
[40] Klein, 3.

A New Legal Regime

A new legal regime must be written with specific clarifying language that enhances the protection of non-combatants and ensures that the burden of war and moral authority is not delegated to LAWS.

The first issue is to clarify language that allows for ambiguity in who or what can carry out IHL obligations. I propose the following additions:

1) "A combatant, commander, or other iterations of responsible parties in IHL shall be a natural person."

2) "Determining if an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, is excessive in relation to the concrete and direct military advantage anticipated shall be the ultimate responsibility of a natural person alone."

3) "In the conduct of military operations, the constant care taken to spare the civilian population, civilians and civilian objects shall be the ultimate responsibility of a natural person.

4) "The decision to kill shall be the sole decision of the responsibility of a natural person alone."

These additions clearly define the role of responsible persons as natural persons in IHL. If a lethal autonomous weapons system cannot be a combatant (attacker), commander (decision-maker), or responsible party, then it must be considered an instrument of warfare. Instruments do not possess moral authority, legal status, and cannot carry out IHL obligations that require that agency or status. While it does not necessarily prevent LAWS from performing *any* actions in war that may contribute in part or in full of fulfilling IHL obligations, it does require human control in the areas that have inherent moral burdens.

This responsibility should not be considered burdensome human oversight. Not only are there multiple ways to carry out IHL obligations, but because humans are involved in the decision making process during war anyway, there are already oversight mechanisms for these obligations

as they exist today. What these additions do is ensure that these specific obligations, and overall responsibility in war are not automated away.

While the above additions delineate areas for sole human control and responsibility, they do not address requirements that should be placed on LAWS for legal use. The following proposed addition establishes the concept of a threshold, although not a specific one:

5) "Any automated technology used in any capacity in the conduct of war must be able to do so with a precision that is at least equal to a reasonable commander."

The upside of this proposal is that it creates a the concept of a threshold using a concept already in use and rooted in human decision-making: a reasonable commander. The glaring downside with any threshold question is that it will be subject to broad interpretation.

A potential remedy might be to define a reasonable commander in terms of the knowledge they should have about LAWS.

6) "A reasonable commander who has in their charge an autonomous system must be aware of the capabilities, limitations, and impacts of potential failures that may be exacerbated by its design and must take these into consideration when evaluating their responsibilities under International Humanitarian Law."

This proposal makes clear the expectations of a reasonable commander. Like the other proposal, it does not specifically express the technical knowledge required. Still, it places the burden of understanding AI technology on the commander, effectively removing any potential arguments shifting blame from the responsible commander to the AI because of technical ignorance. This proposal forces countries to consider the caliber of commander they want to place in charge of LAWS. Its follow-on effect is that by placing a high level of responsibility on commanders, which is nothing new, countries will be more considerate in developing, purchasing, and employing systems to protect their militaries from criminal liability.

None of these proposals should be read to constrict the development of new technology,

but rather accommodate the primary goal of artificial intelligence in war; that is, to optimize armed forces in the information age. Further, they are easy to adhere to, and militaries can accomplish them in many ways. The proposals address the ambiguity and concerns surrounding moral obligations and IHL responsibilities and leave countries free to decide *how* they want to accomplish them and the moral and legal risks they are willing to accept in ignoring them.

III
CRIMINAL RESPONSIBILITY

After establishing the borders of Lethal Autonomous Weapons Systems' moral and legal agency, we find ourselves dealing with the problem of accountability. If an autonomous system commits a war crime, who is responsible? Is it the first human in the LAWS chain of command? The state? One or more of the engineers who designed it? There may be a place for criminal or civil liability for each of these parties, but this paper will focus on criminal liability for the immediate humans in a "command" or "supervisory" position over the LAWS.

The answer to the question of whether "humans are criminally liable for the actions of a LAWS" is in the affirmative. LAWS are tools with no moral or legal agency and do not have legal personalities. Because of this lack of agency and the status of LAWS as instruments of war, legal obligations and responsibilities cannot be attached to autonomous weapons, no more than they can be attached to a howitzer.

Attributing criminal acts perpetrated by LAWS to a human presents a similar problem as the delineation of IHL responsibilities. Namely, there is no explicit language in any statutes or treaties dealing with the issue of autonomous weapons. While the International Criminal Court (ICC) does use terms such as "person" or "natural person," the concept of responsibility becomes muddled when the *actus reus* is caused by a decision stemming from an autonomous machine.[41] While LAWS may have no moral or legal agency, they can still "commit" a war crime independent of a human order. How, then, is one supposed to connect the crime to the human "commander?"

Compounding this problem is the complexity of AI. A human "commander" may be held liable for a system they do not substantially understand. While this may be a weak excuse for

[41] "Rome Statute of the International Criminal Court" (The International Criminal Court, July 1, 2002).

traditional weapons, it is a reality for LAWS. LAWS will take actions that a human overseeing it may not predict.[42]

COMPLEX TOOLS

When thinking about the problems of LAWS and criminal accountability, it is helpful to use analogies. If a sniper used his rifle to kill a civilian, it would be preposterous to spend time and energy connecting the killing of the civilian to the sniper by way of indirect liability. The sniper is the only perpetrator who physically carried out the prohibited conduct.[43] The rifle did not make any decisions independent of the sniper. The sniper had to load the rifle, set the sights, pick the target, adjust their aim, and apply physical pressure on the trigger before the civilian's death. The rifle is an instrument of war, just like LAWS are. The difference, we know, is that LAWS operate much more independently than a rifle.

Using the Boeing 737 Max crashes as an analogy may fit slightly better. In brief, authorities attributed the cause of the crashes to issues with the Maneuvering Characteristic Augmentation System (MCAS).[44] This flight control law system receives inputs from aircraft sensors, makes adjustments to the aircraft's controls, and compensates the plane if necessary.[45] While this system does not involve machine learning, it is algorithmically based and takes actions independent of a human.

If we used the MCAS as a stand-in for LAWS and the deaths from the crashes as a war crime, could the pilot be held responsible? While the pilot did not design the system, they were

[42] I will cast aside some differences between the ICC, other tribunals, and common and civil law courts when discussing *mens rea* standards and other elements of crime. Acknowledging that these standards are important in prosecution, my primary purpose is not under what circumstances and jurisdictions a human controller of AI may be prosecuted but the shortcomings in general principles surrounding modes of liability, knowledge, and *mens rea*.
[43] Antonio Cassese, *Cassese's International Criminal Law* (Oxford University Press, 2013), 161.
[44] "Joint Authorities Technical Review: Observations, Findings, and Recommendations," October 2019.
[45] "Joint Authorities Technical Review: Observations, Findings, and Recommendations."

operating it. Whether or not the pilot had a robust understanding of the MCAS doesn't necessarily matter if it is assumed that the pilot *should have* had a robust understanding of the system. Regardless of the pilot's knowledge, the MCAS is a tool or instrument of flight. The tool may operate in an unforeseen way, but this does not make the tool criminally liable. Somebody else must shoulder this liability. While not as absurd as the rifle analogy, it may seem odd to connect the pilots of the planes to the deaths of the passengers by way of indirect liability, at least without understanding the degree to which the pilots were trained on the system.

Moving out of analogies and into discussions on artificial intelligence adds another layer of complexity. An AI-enabled system has even greater autonomy than the MCAS. Like the MCAS and the rifle, this system is an instrument. However, this tool of war differs from the sniper rifle in that it can substitute human inputs and decisions for its own. It differs from the MCAS because, while the MCAS is independent of human control, LAWS "understand higher-level intent and direction."[46] This positions LAWS to be a tool, yet also a non-human direct perpetrator of a crime.

THE KNOWLEDGE PROBLEM

In evaluating a commander of an artillery unit, it is assumed that they understand the equipment they are using, but during prosecution it does not always play out so simply. Even simple weapons have issues surrounding their capabilities and limitations. In *Gotovina et al.*, there were serious discrepancies regarding accepted accuracy ranges with artillery pieces—weapons that are not technologically sophisticated.[47]

The Appeals Chamber in *Gotovina* wrote that the Trial Chamber erred in establishing its "impact analysis," not only because of divergent opinions of error ranges between subject matter

[46] Vincent Boulanin and Maaike Verbruggen, "Mapping the Development of Autonomy in Weapons Systems" (Stockholm International Peace Research Institute, November 2017).
[47] Gotovina et al. (IT-06-90) "Operation Storm," No. IT-06-90-A (International Criminal Tribunal for the former Yugoslavia November 16, 2012).

experts, but also because of additional disregarded for testimony around external "factors such as wind speed that would affect range of error."[48] The faults in this "impact analysis" proved to be significant enough and crucial enough to the rest of the case that it was able to "undermine the Trial Chamber's conclusion that artillery attacks…were unlawful."[49]

The more complicated capabilities of the MCAS were misunderstood in a different manner than *Gotovina.* In this case, the system's complexity was not only misunderstood by the operators but also by the regulating authority, the Federal Aviation Administration (FAA). The after-action report stated:

> "The FAA was not completely unaware of MCAS; however, because the information and discussions about MCAS were so fragmented and were delivered to disconnected groups within the process, it was difficult to recognize the impacts and implications of this system. If the FAA technical staff had been fully aware of the details of the MCAS function…the agency likely would have required an issue paper for using the stabilizer in a way that it had not previously been used."[50]

This statement is two warnings wrapped into one: The first is that the regulators did not understand the technology and did not require pilots to be trained appropriately. Autonomous systems often have layers and segments of technology developed separately and have different machine learning processes that require different evaluation metrics.

The second warning is that the bureaucratic process in place was not equipped to handle its complexity.[51] In relating this case to this paper, the bureaucratic process can be read as a "legal process." Both of these failures combined to create the environment where an algorithmic-based system failed, and the human operator did not know how to take action to prevent the failure.

---

[48] Gotovina et al. (IT-06-90) "Operation Storm" paragraph 60.
[49] Gotovina et al. (IT-06-90) "Operation Storm" paragraph 83.
[50] "Joint Authorities Technical Review: Observations, Findings, and Recommendations," 13.
[51] The misunderstanding of MCAS was at least partially because of intentional misleading by Boeing. This fact should be taken as a warning for emerging technology but does not affect the lessons drawn out in this paper.

The MCAS scenario presents a different problem but not an insurmountable one. The complexity of the technology led to the failure, but, had better procedures been in place, a trained pilot could have taken corrective action to right the plane, even if they did not understand the MCAS at a granular level.

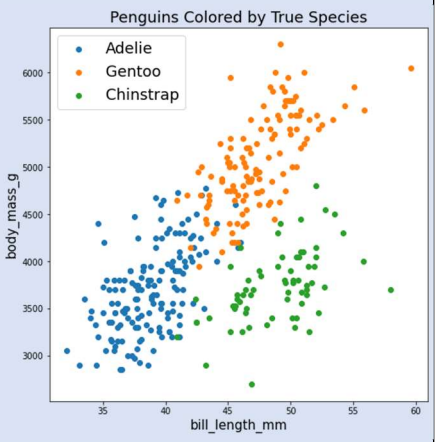
LAWS, on the other hand, presents a knowledge feasibility issue. Some of the components of AI can be understood with tech manuals, and others are simply "black boxes." Others still contain such a vast amount of information that it is infeasible for a human to understand every

DATA IN AI

The use of data is the most widely recognized impactable aspect of machine learning. Suppose one aspect of LAWS is the automated targeting of enemy military vehicles, and the machine learning system is trained on Russian tanks exclusively. In that case, the system will have a high error rate if deployed in the Middle East. The system will not target hostile vehicles in this scenario because it doesn't recognize them as threats.

The more nuanced the examples, the more difficult it is for the machine to make determinations and the human to understand how or why an autonomous system behaves the way it does. An innocuous example shown in the diagram below of penguin species grouped by body mass and bill length illustrates this point.

If a system uses the data visualized on the graph and tries to identify a Gentoo penguin, it will be easy if its characteristics fall in the center of the mass of orange dots. The denser the grouping of the dots, the more consistent the input data, and, therefore, the higher the accuracy rates. However, if the Gentoo penguin's characteristics fall closer to the center, overlapping with the other species, accuracy declines.



Input data and accuracy rates are just two critical aspects that complicate machine learning. Model type is also crucial. Model types, such as linear regression or clustering, determine how the machine draws inferences between what features it prioritizes as important.

The rules that govern what features to use are called parameters and hyper-parameters. If the correct model is not chosen alongside appropriate hyperparameters, the machine may choose completely unrelated features to prioritize and then prioritize them in an unfavorable way. For example, using hairstyle to determine combatant status.

What makes this an even more drastic problem is that a human may not know that this correlation is happening until the system misidentifies a target.

aspect. If the law requires a certain threshold of knowledge in criminal liability for an autonomous weapon failure that results in a crime, it must reckon with what is an appropriate amount of knowledge for a LAWS supervisor to possess and what amount of knowledge would make deploying the weapon impossible.

SCENARIO

The following scenario is constructed from various international criminal cases and news stories on artificial intelligence in war.[52] *It will be used to evaluate modes of liability and mens rea standards.*

During an International Armed Conflict (IAC), a human military pilot of the rank of Lieutenant Colonel (O-5) is in command of an autonomous drone squadron.[53] The pilot is flying a routine mission in the vicinity of a town of moderate size.

The pilot is in the passenger seat of a plane flown by AI. The pilot's primary mission is to provide early warning and air defense of their base by commanding the autonomous drone squadron. Early warning and air defense is accomplished by identifying, evaluating, and intercepting (if necessary) inbound air contacts. The squadron stays within the vicinity of the pilot's plane unless tasked by the pilot to go elsewhere.

The pilot's secondary mission is to conduct airstrikes on targets provided via communication uplink from the pilot's base during the flight. These strikes are carried out by individual or multiple drones in the squadron, depending on the target.

The AI squadron has three modes:

1. Non-autonomous (drones fly in an idle pattern until given orders). The drones will only

[52] Both the NATO bombing of Albanian refugees near Gjakova, and the shelling of a soccer stadium of Dobrinja were drawn on for facts in this scenario to make it as realistic as possible. Information on the drone swarm and AI pilots are cited when appropriate.
[53] Sue Halpern, "The Rise of A.I. Fighter Pilots," *The New Yorker*, January 17, 2022, https://www.newyorker.com/magazine/2022/01/24/the-rise-of-ai-fighter-pilots.

respond to a command given by the pilot in this mode. They will not respond even if fired upon.

2. Semi-autonomous or "man in the middle." The drones will patrol in patterns and intercept targets identified as hostile by a human but do not conduct offensive operations without human approval. The drones will conduct airstrikes only with approval from the pilot.

3. Fully autonomous: The human pilot can only negate the AI's actions. The drones will assess and neutralize any threats in the air or on the ground, entirely independent of human control. The drone squadron may conduct airstrikes on uplinked targets if the drone evaluates it as appropriate use of resources.

For our scenario, we assume that the autonomous weapons system was operating as designed; there were no perfidious actions by any party.

ACTUS REUS

The pilot takes off with the drone squadron, reaches their patrol point, and sets the AI squadron in "semi-autonomous" mode. The pilot then begins to screen recommended targets uplinked from base and allocates drones to conduct strikes. Every action thus far is in full compliance with International Law.

As the mission progresses, a sudden increase in radar contacts coincides with an intelligence report from base that the enemy is launching an air attack alongside ground-supported air defense batteries. These two developments cause the pilot to become overwhelmed with hostiles, and they begin to fear being shot down. In response to a sudden and drastic increase of enemy forces, the pilot sets the AI squadron to "fully-autonomous" to deal with incoming threats.

The AI squadron identifies and evaluates humans on the ground and classifies them as a threat. These turn out to be primarily non-combatants. It injures 130–140 civilians and kills 14 civilians in an attempt to suppress enemy air defense.

The pilot and squadron return to base, and charges are considered for the illegal act of the indiscriminate targeting of civilians.

MENS REA

This section will outline different mental elements. Two main pieces of information will be vital in establishing what type of *mens rea* the pilot had. The first is the knowledge of civilians in the area of the pilot's mission. We assume the pilot had full knowledge that there was a high probability, if not certainty, of non-combatants in the area. The pilot was of sufficient rank and, therefore experience, to fully comprehend their mission. Additionally, we assume that the pilot was familiar with the surrounding area through pre-flight briefings.

The second piece of information will be explored further: if the pilot knew, and if so, to what extent, of the consequence that the autonomous drone squadron would kill non-combatants. To use Finnin's definition, "knowledge means awareness that a circumstance exists or consequence will occur in the ordinary course of events."[54]

This evaluation will set aside "willful blindness."[55] Also, there is room in this scenario for excuses or justifications because of extenuating circumstances. However, these are subjects for a separate discussion and will not be considered here.

Direct Intent

For the pilot to have direct intent, they would have to want, or have acted purposefully, to target and kill non-combatants. This is not the case in this scenario. The pilot activated the fully autonomous mode to save their own life, not to target civilians intentionally. However, even if this were the case, there does not need to be a high cognitive threshold for direct intent.[56] We can leave this classification as not germane to our scenario.

---

[54] Sarah Finnin, "Mental Elements Under Article 30 of the Rome Statute of the International Criminal Court: A Comparative Analysis," *The International and Comparative Law Quarterly* 61, no. 2 (April 2012): 326.
[55] Finnin, "Mental Elements Under Article 30 of the Rome Statute of the International Criminal Court: A Comparative Analysis."
[56] Finnin, 330.

Oblique Intent

Oblique intent is the first realistic gradation we can apply to our pilot. For oblique intent to apply, the pilot would have to have "certain," "highly probable," or "virtual certainty" knowledge about the consequences.[57] This places the knowledge threshold quite high for the pilot. Given the pilot's rank, experience, and the fact that they were nominally in control of this squadron, it is foreseeable that the pilot could be seen to have oblique intent.

Which of the components of AI would the pilot need to understand to be certain the autonomous system would target civilians? How deep does this knowledge need to be? Understanding the real time input data and the training data *vis-à-vis* the pilot's circumstances would be necessary to reach this level of intent. If the training data and real time input data were different enough to cause the drone squadron to have decreased and therefore dangerous inaccuracy, and the pilot knew this, "virtual certainty" *may* be an attainable level of knowledge to prove.

However, what level of this disparity between training data and the scenario would matter as well as how that disparity impacted the AI's actions. Measuring this disparity and its effect on the crime would probably be very difficult. Returning to a previous example, if the LAWS were trained to identify Russian tanks, but then used in the Middle East, this may be a large and measurable disparity. But what if the LAWS were trained on Russian tanks operating in a desert, and then deployed to a country using Chinese tanks in a desert environment? Assuming the disparity is this clear, how then will the disparity be measured to establish intent? Assuming the LAWS was at least trained on similar data in a similar environment, "virtual certainty" seems unattainable.

[57] Finnin, 332.

In addition to understanding the training data, accuracy rates would most likely need to be known. If the pilot knew that the AI performed poorly in circumstances similar to the pilot's, then the knowledge threshold would be reached. However, if the machine learning performed with high enough accuracy, this, in addition to understanding the training data, wouldn't be enough to reach the high cognitive threshold.

Knowing and understanding which models are used—and the models' parameters in conjunction with data and accuracy rates—may arguably be enough to rise to "virtual certainty." However, this component begins to enter what is feasible for the pilot to understand as it pertains to a specific circumstance. Say the pilot did understand how the LAWS made inferences and grouped data, it is unlikely the pilot could foresee how this would impact the AI's decision-making in the specific scenario described. To link knowledge of training data, models, and parameters to a high level of certainty, a prosecutor would need to evaluate analogous scenarios during the pilot's training or other missions they flew in with the same AI.[58] This training scenario should be enough to make the case that the pilot understood *similar* situations and should have been able to foresee similar events unfolding.

Recklessness

The pilot almost certainly could be said to have the lower level of reckless intent. This would mean that the pilot is not certain that the civilians will be killed but is aware of the risk that it may occur.[59] Cursory knowledge of the training data, alongside knowledge of the models, should be enough to reach this "conscious risk taking."[60]

[58] If the AI squadron in the scenario was consistently "learning" during combat missions, there may be technical limitations to the gathering this evidence forensically. If the squadron was not "learning" during combat missions, it may be appropriate to look at other types of AI the pilot interacted with to establish a level of knowledge.
[59] Finnin, "Mental Elements Under Article 30 of the Rome Statute of the International Criminal Court: A Comparative Analysis," 333.
[60] Finnin, 334.

Accuracy rates alone may elevate the pilot to being aware of the risk. For example, if the pilot knew that accuracy rates declined when the number of contacts increased, the pilot would have understood the danger of misclassification when placing the weapon into fully autonomous mode because of increased contacts.

Culpable or Gross Negligence

Negligence requires a person to "abide by certain standards of conduct or take certain specific precautions with which any reasonable person should comply,"[61] but the "certain standards of conduct" are unwritten. What the standard is, or what a reasonable person should do, given that there has yet to be any person in charge of an AI weapon, may make this level of intent unreachable out of the sheer novelty of LAWS. As LAWS are developed, incorporated into different aspects of warfare, and it becomes clearer what standards of conduct might be, negligence may become a viable option. In our scenario, it may be too low given that recklessness would fit the level of knowledge a pilot of the rank described would already have.

MODES OF LIABILITY

In establishing that the pilot is not the principal perpetrator of the crime, but not entirely in control of the autonomous weapon, leaves only indirect liability for criminal accountability. To highlight the various modes of liability in international law and their shortcomings, I will discuss Aiding and Abetting, Joint Criminal Enterprise (JCE), and Command Responsibility. Ordering, Instigating, Planning, and Residual Accessory all have unique problems, but the main issues of the three I will discuss are sufficient to derail the others.

Aiding and Abetting

Aiding and abetting in our context would require that a third party give practical assistance

[61] Cassese, *Cassese's International Criminal Law*, 53.

to the perpetrator.[62] In our case, we would need to read "practical assistance" as deploying LAWS or placing LAWS in a specific autonomous setting. We would also have to accept that a system acts in the role of "perpetrator," a role understood as designated for humans with a legal personality. To accept that an object can perpetrate a crime would run in direct contradiction to the idea that LAWS are instruments without moral or legal agency. There is certainly a level of independence in autonomous weapons that is not present in other weapons systems, but that does not alter LAWS status as an object devoid of legal status.

The "object-as-perpetrator" issue is enough to make this mode of liability insufficient for use in prosecuting the pilot. Even ignoring this issue, the only way the pilot could "aid" the LAWS would be to place the LAWS in a particular setting, which would then make the pilot the active perpetrator of the crime and thus not in need of a mode of liability.

Joint Criminal Enterprise

JCE requires that "the prosecution prove i) the involvement of a plurality of persons in the commission of the crime; ii) the existence of a common plan, design, or purpose which amounts to or involves the commission of a crime; and iii) the participation of the accused in the JCE in the form of a significant contribution."[63]

For JCE to work, one would have to consider the LAWS and the human overseeing it as a "plurality of persons." This is already a stretch. Assuming we read persons and autonomous weapons into criteria "i," the next issue is that "ii" and "iii" mean the LAWS would be "consciously" disregarding international law and possibly disregarding standing orders from its operator or programming. At this point in LAWS development, a scenario such as this delves too far into science fiction to be discussed in this paper. Even if the human operator and the LAWS

[62] Cassese, 193.
[63] Cassese, 163.

had a common plan to commit a war crime, the significant contribution of the human would be to deploy the LAWS. Just as in aiding and abetting, this would turn the pilot into the active perpetrator of the crime and no mode of liability is required.

JCE i, ii, and iii all have significant problems along the lines just outlined, and just like the other modes of liability, it requires taking substantial departures from the instrumentalist position on AI. Blurring the lines between weapon and human and calling a LAWS "perpetrator" begins to remove its status as an instrument of war.

Command Responsibility

According to Article 28 of the Rome Statute, command responsibility requires that a military commander has control over their subordinates, has knowledge, or constructive knowledge, of their crimes, fails to prevent or punish these crimes, and the crimes are a result of the superior officer's failings.[64]

Different from the other modes of liability, command responsibility uses not the word "persons" but "forces."[65] "Forces" in noun form is defined as "persons or things."[66] Admittedly, this was probably not the framers' intention, yet it could open the door for applying this mode to AI. This notwithstanding, the same problem of "object-as-perpetrator" arises, regardless of the language.

The knowledge aspect is rife with the problems discussed earlier in the paper, but from a prosecutor's point of view, it is not unprovable, especially because of the lower *mens rea* standard found in the phrase "should have known."[67] Linking the superior's failure to the crime would go hand in hand with proving the pilot's level of knowledge.

[64] Cassese, 187.
[65] "Rome Statute of the International Criminal Court," sec. 28.
[66] *Merriam-Webster Dictionary*, accessed May 4, 2022, https://www.merriam-webster.com/dictionary/forces.
[67] "Rome Statute of the International Criminal Court," sec. 28.

Where command responsibility falls apart is "effective command and control, or effective authority and control."[68] The whole point of autonomous weapons is that they can operate independently of a human controller. Similar to subordinate human troops, a superior officer is not in a place to micromanage a private on patrol any more than an autonomous weapon. However, the striking difference is that the tools at the commander's disposal to ensure a human's compliance with orders is not the same as an autonomous weapon.

First, a commander can punish a subordinate. This punishment works as a deterrence. No such tool is available to "punish" a LAWS and deter insubordination. Second, the commander can rely on societal norms to ensure compliance. A subordinate troop perceives the environment differently from a machine in that humans have moral agency and moral perception. There is a possibility that a human will not act because of moral misgivings. There is no such possibility for a machine because there is no societal or moral reputation at stake for an autonomous weapon as there is for a human. Similarly, a subordinate troop is subordinate not just to the highest-ranking officer but also to every higher-ranking superior up and down the chain of command and then informally to his peers. A LAWS has no peers and no recognition of a "chain of command."

Effective control is further challenged because of the speed and complexity of the weapon system. If it is unfeasible for a commander to understand every complex machine learning component, how can a commander exercise effective control over it?

A New Mode: Supervisory Responsibility

Two problems—the complexity of AI complicating the establishment of a knowledge threshold and an object seemingly being a perpetrator of a crime yet having no legal personality—are not easy to solve. The solution may be legal gymnastics in existing modes of liability or *mens*

[68] "Rome Statute of the International Criminal Court," sec. 28.

rea standards, but this type of problem-solving is not conducive to fixing the unique problem LAWS present. The best, and most sure, way is a paradigm shift in how we think about legal responsibility and principal perpetrators. New standards must follow.

A new mode of liability should be constructed specifically for autonomous weapons. It could be labeled as "supervisory responsibility" to indicate less effective control but still a position of responsibility over the tool. This new mode would also need to use language specifically for autonomous systems. This language and mode of liability would indicate that the object is the principal "perpetrator" of the crime but automatically links the crime to the human supervisor. This acknowledges that the action was outside the human's control, yet they were in a position where they *should have* foreseen the consequences arising.

The problem of establishing a knowledge threshold is much more challenging. Part of the difficulty is in how this threshold would be laid out in either treaty, domestic regulation influencing customary law, or maybe an addendum to the ICC statute.

The second difficulty is each type of autonomous technology is different. The example I used to look at various AI components was specific to visual identification. While the principles are the same, automated driving, nonvisual identification, stability control, munitions selection, and many other AI applications will all have different intricacies in how they make decisions. More importantly, machine learning technology is currently not at a point where humans can even understand the decision-making process.

Despite the complexity, there are nonetheless core parts of AI that could make a good starting point for required knowledge. Training data, accuracy rates, type of model and parameters, and failure scenarios should be sufficient to allow a human supervisor to make informed decisions about how they use their autonomous weapon.

IV
THE THRESHOLD QUESTION

There are three "threshold" questions that autonomous weapons pose. The first was discussed in the first section of this paper and asks if there is an inherent moral burden in an IHL obligation that makes an autonomous weapon morally unable to carry it out.

The following section addresses the next two threshold questions, the first of which probes: What accuracy rate should an autonomous weapon have, or how much uncertainty and potential for a mistake are we willing to tolerate in a weapon? The second question addresses the threshold of knowledge for a human supervisor. Since it may be impossible for a human to reasonably understand AI in the same capacity that an artillery officer may understand a howitzer, what level of knowledge should we expect?

Mathematical certainty in the tasks LAWS may be asked to do, such as the drone squadron in our scenario, may never be reached. In the law of war, LAWS accuracy could be characterized as "doubt." Assuming we can morally reconcile this idea of an autonomous weapon being deployed short of being 100% accurate, what *technical* rate of accuracy (or conversely, what rate of failure) is acceptable? 99%? 89%? Once a number is decided, how would that work with IHL, if "in a case of doubt whether a person is a civilian, that person shall be considered a civilian."[69]

The only way this could work is if a quantifiable accuracy rate were agreed upon. This assumes, however, that accuracy measurement can be standardized across machine learning platforms and completely ignores the pragmatic problem of nations being unwilling to turn over these accuracy rates to anyone for any reason. Perhaps a real-time accuracy rate could be projected to a human supervisor, and this could play into the supervisors' decision on how to utilize the AI.

[69] Protocol Additions to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (AP/I), art. 50.

This may avoid having to declare a set accuracy rate across all AI platforms and could be tailored to a circumstance under the human guidance of a supervisor in command. A "spectrum" of accuracy may be practical, but if speed is the driving reason behind using LAWS, then forcing a human in the loop at every juncture makes the technology impractical to use.

The knowledge threshold question is equally as complicated as the accuracy one. This threshold is twofold: first, how much knowledge should a supervisor have to operate the weapon system, and second, how much knowledge will make them culpable for a crime? It may appear that the threshold should be the same, but it is not so easy. Using artillery, for example, an officer may be well versed in the munition defect rate, say 1 in every 10,000 rounds will explode prior to impacting the ground. Should that officer be held accountable when that round explodes out of specification and kills a civilian? Just like the more complicated AI systems, the artillery officer doesn't know which of the 10,000 rounds is defective, only that *statistically* one of them is. His knowledge, therefore, is different in what is required to operate the weapon versus what is needed to hold him criminally liable.

A problem present in AI but not in the artillery analogy is the number of components involved in a weapons system. At what level each of these components needs to be (or *should be*) understood is a separate question. On one extreme would be an extremely high threshold where the supervisor must be intimately familiar with every aspect of the machine learning process. One could imagine requiring a doctorate or training similar to an astronaut in order to supervise these weapons. The obvious problem with this threshold is that it makes the technology near impossible to deploy.

The other side would be an extremely low threshold, where a supervisor is little more than an operator who powers on and off the system. The problem in this scenario, other than the ethical

problems of placing unknowledgeable people in charge of complex systems, is it opens up almost anyone to a command responsibility style of liability. This would defeat the purpose of a crime predicated on positional authority and responsibility in wartime.

The middle ground is a partial understanding threshold. This would entail understanding training data, whether it was skewed toward a particular region, and understanding the ramifications of using the weapon in a different area. This knowledge of how the system works is present in this threshold, as is the ability that consequences *could* arise, but those exact consequences are unknown.

CONCLUSION: MORAL RISK

In using autonomous weapons systems, there are three paths the world can take. The first is an outright ban, while the last is a *laisse-faire* approach. I do not think either of these is prudent.

The middle approach, one characterized by regulated use of LAWS, has its own problems. This approach must address and answer the threshold questions previously discussed, as well as adopt a new IHL regime and criminal mode of liability. More importantly, however, allowing LAWS to operate on the battlefield will signal acceptance of collateral damage to non-combatants within the acceptable failure rate of AI weapons and openness to the possibility of prosecuting combatants who only partially understand the AI.

The moral risk of this approach should be evident. Unlike the munition that malfunctions and explodes killing a few civilians, a mishap by an autonomous weapon could quickly mount a high body count. It also introduces risk into a concept that many countries do not accept: a machine deciding to kill a human. If LAWS are to be deployed, there must be an acceptance of some amount of moral risk that a machine, being imperfectly accurate, will kill the wrong target and be unburdened by this result.

This paper focused exclusively on Lethal Autonomous Weapons Systems that were designed to engage in the IHL decision cycle and ultimately kill a human. In outlining the ethical concerns with such systems, it offered up a legal framework that answered these problems. It summarized problems with accountability if the AI and the new IHL framework were to fail, resulting in a war crime, and offered up a new mode of liability. Finally, in addressing the pragmatic portion of introducing AI weapons onto the battlefield, it asserted that the international community would have to accept moral risk in the technical and pragmatic realities of AI accuracy and supervisory knowledge.

There are other options, however. The natural evolution for AI does not have to fully automate the targeting cycle independent of human involvement. For the time being, AI can be regulated to Staff Corps type positions conducting tasks as administrative work, supply chain routing, parts ordering, munitions selection, or "first look" contact identification. This may not be the science fiction AI war of the future, but these are areas where battles and war are won and lost.

BIBLIOGRAPHY

Bellaby, Ross W. "Can AI Weapons Make Ethical Decisions?" *Criminal Justice Ethics* 40, no. 2 (May 4, 2021): 86–107. https://doi.org/10.1080/0731129X.2021.1951459.

Bergman, Ronen, and Farnaz Fassihi. "The Scientist and the A.I.-Assisted, Remote-Control Killing Machine." *The New York Times*, September 18, 2021, sec. World. https://www.nytimes.com/2021/09/18/world/middleeast/iran-nuclear-fakhrizadeh-assassination-israel.html.

"Boeing: The 737 MAX MCAS Software Enhancement." Accessed May 2, 2022. https://www.boeing.com/commercial/737max/737-max-software-updates.page.

Boulanin, Vincent, Netta Goussac, and Laura Bruun. "Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human–Machine Interaction." SIPRI, June 2021.

Boulanin, Vincent, and Maaike Verbruggen. "Mapping the Development of Autonomy in Weapons Systems." Stockholm International Peace Research Institute, November 2017.

Cassese, Antonio. *Cassese's International Criminal Law*. Oxford University Press, 2013.

de Jongh. "Statement of the Netherlands." Statement presented at the Group of Governmental Experts on LAWS, Geneva, April 26, 2019.

"Deep Learning vs. Machine Learning - Azure Machine Learning," January 20, 2023. https://learn.microsoft.com/en-us/azure/machine-learning/concept-deep-learning-vs-machine-learning.

Dempsey, Jim. "Managing the Cybersecurity Vulnerabilities of Artificial Intelligence." *Lawfare* (blog), November 17, 2021. https://www.lawfareblog.com/managing-cybersecurity-vulnerabilities-artificial-intelligence.

"DHP-D258-Classification_and_Regression_with_KNN.Ipynb." Accessed May 3, 2022. https://colab.research.google.com/drive/1PTFpYeNFbyUY0Ig_gXrpADjRaEnXiL3A?usp=sharing.

Dinstein, Yoram. *The Conduct of Hostilities Under the Law of International Armed Conflict*. Cambridge University Press, 2004.

"Directive on Automated Decision-Making." Government of Canada, February 5, 2019. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

Finnin, Sarah. "Mental Elements Under Article 30 of the Rome Statute of the International Criminal Court: A Comparative Analysis." *The International and Comparative Law Quarterly* 61, no. 2 (April 2012): 325–59.

Gotovina et al. (IT-06-90) "Operation Storm," No. IT-06-90-A (International Criminal Tribunal for the former Yugoslavia November 16, 2012).

Halpern, Sue. “The Rise of A.I. Fighter Pilots.” *The New Yorker*, January 17, 2022. https://www.newyorker.com/magazine/2022/01/24/the-rise-of-ai-fighter-pilots.

Heller, Kevin Jon. “‘One Hell of a Killing Machine’: Signature Strikes and International Law.” *Journal of International Criminal Justice* 11, no. 1 (March 1, 2013): 89–119. https://doi.org/10.1093/jicj/mqs093.

“Joint Authorities Technical Review: Observations, Findings, and Recommendations,” October 2019.

Kahneman, Daniel. *Thinking Fast and Slow*. Farrar, Straus and Giroux, 2013.

Klein, Gary. *Sources of Power: How People Make Decisions*. MIT Press, 1999.

Lewis, Dustin A. “Three Pathways to Secure Greater Respect for International Law Concerning War Algorithms,” 2020, 33.

Lewis, Dustin A., Gabriella Blum, and Naz K. Modirzadeh. “War-Algorithm Accountability.” Harvard Law School Program on International Law and Armed Conflict, August 2016.

Merriam-Webster Dictionary. Accessed May 4, 2022. https://www.merriam-webster.com/dictionary/forces.

Nasu, Hitoshi. “The Kargu-2 Autonomous Attack Drone: Legal & Ethical Dimensions.” *Lieber Institute West Point*, June 10, 2021. https://lieber.westpoint.edu/kargu-2-autonomous-attack-drone-legal-ethical/.

Protocol Additions to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts (Protocol I) (AP/I) (1977).

“Rome Statute of the International Criminal Court.” The International Criminal Court, July 1, 2002.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Third. Prentice Hall, 2009.

Sayler, Kelley M. “Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems.” Congressional Research Service, n.d. https://crsreports.congress.gov.

Schmitt, Michael N. “Autonomous Weapons Systems and International Humanitarian Law: A Reply to the Critics.” *Harvard Law School National Security Journal* Journal 4 (2013): 1–37.

Schwarz, Elke. “Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control.” *Philosophical Journal of Conflict and Violence* 5, no. 1 (May 20, 2021): 53–72. https://doi.org/10.22618/TP.PJCV.20215.1.139004.

Trabucco, Lena, and Kevin Jon Heller. “Beyond the Ban: Comparing the Ability of ‘Killer Robots’ and Human Soldiers to Comply with IHL.” *Fletcher Forum of World Affairs* 46, no. 15 (April 21, 2022). https://ssrn.com/abstract=4089315.

United States Department of Defense. Autonomy in Weapon Systems, 3000.09 § (2012).

Woods, Andrew Keane. “Robophobia.” *University of Colorado Law Review*, no. 1 (March 17, 2022).

www.javatpoint.com. “Hyperparameters in Machine Learning - Javatpoint.” Accessed May 3, 2022. https://www.javatpoint.com/hyperparameters-in-machine-learning.

Zewe, Adam. “These Neural Networks Know What They’re Doing.” *MIT News | Massachusetts Institute of Technology*. Accessed November 13, 2021. https://news.mit.edu/2021/cause-effect-neural-networks-1014.