[Accepted for publication in Journal of National Security Law & Policy, forthcoming 2018 (pre-edited version)

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND THE USE OF FORCE BY STATES

Ashley Deeks[*], *Noam Lubell*[**] *& Daragh Murray*[***]

"[O]ne bad algorithm and you're at war."[1]

I. INTRODUCTION

Big data technology and machine learning techniques play a growing role across all areas of modern society. Machine learning provides the ability to predict likely future outcomes, to calculate risks between competing choices, to make sense of vast amounts of data at speed, and to draw insights from data that would be otherwise invisible to human analysts.[2] As such, the use of machine learning presents a significant opportunity to transform how we understand and engage with issues across a range of subject areas, and to use this more-developed understanding to enhance decision-making. Although still at a relatively early stage, the deployment of machine learning algorithms has already begun. For example, doctors use machine learning algorithms to match symptoms to a particular illness, or to

[*] Of the University of Virginia Law School

[**] University of Essex

[***] University of Essex. This work was supported by the Economic and Social Research Council grant number ES/M010236/1.

[1] Jenna McLaughlin, *Artificial Intelligence Will Put Spies Out of Work*, FOREIGN POLICY, June 9, 2017, (quoting Justin Cleveland).

[2] LEE RAINIE & JANNA ANDERSON, PEW RESEARCH CENTER, CODE-DEPENDENT: PROS AND CONS OF THE ALGORITHM AGE 2-41 (2017),

identify individuals at risk of developing a particular health condition to target preventive intervention.[3] In a criminal justice context, courts use machine learning algorithms to establish individual risk profiles and to predict the likelihood that a particular individual will re-offend.[4] The private sector uses similar techniques for tasks as diverse as determining credit ratings and targeting advertising.[5]

The advantages that flow from machine learning algorithms mean that it is inevitable that governments will begin to employ them—in one form or another—to help officials decide whether, when, and how to resort to force internationally. In some cases, these algorithms may lead to more accurate and defensible uses of force than we see today; in other cases, states may intentionally abuse these algorithms to engage in acts of aggression, or unintentionally misuse algorithms in ways that lead them to make inferior decisions relating to force. Indeed, it is likely that machine learning techniques are already informing use of force decisions to a greater degree than we appreciate. For many years, intelligence agencies have steadily increased their use of machine learning.[6] Although these current uses are likely to be several steps removed from decision-making at the national leadership level,

[3] *See, e.g.*, Daniel Fagella, *7 Applications of Machine Learning in Pharma and Medicine*, TECHEMERGENCE, Jan. 11, 2018.

[4] *See* Wisconsin v. Loomis, 2016 WI 68.

[5] *See, e.g.*, Bill Hardekopf, *Your Social Media Posts May Soon Affect Your Credit Score*, FORBES, Oct. 23, 2015; Aaron Rieke, *Google was right to get tough on payday loan ads – and now, others should follow suit*, MEDIUM, May 13, 2016.

[6] *See, e.g.*, Office of the Director of National Intelligence, *Analysis: Current Research* (listing current research projects, which include a variety of machine learning, artificial intelligence, and big data efforts); David Anderson Q.C., Report of the Bulk Powers Review, para. 8.31(d), Aug. 2016 (quoting MI5 paper stating that the "rapid development of new technologies and data types (e.g. increased automation, machine learning, predictive analytics) hold additional promise").

intelligence analysis heavily influences use of force decisions.[7] It is likely that machine learning algorithms increasingly will inform this analysis. Despite the significant attention given to machine learning generally in academic writing and public discourse, there has been little analysis of how it may affect war-making decisions, and even less analysis from an international law perspective.

Of course, one should not simply assume that machine learning will give rise to new legal considerations. Indeed, issues relating to new technology may often be resolved by the relatively straightforward application of existing law. With respect to the law on the resort to force, however, there are three areas in which machine learning either raises new challenges or adds significant complexity to existing debates. First, the use of machine learning is likely to lead to the automation of at least some types of self-defense responses. Although some militaries currently deploy automated response systems, these tend to be fixed, defensive systems designed only to react to incoming threats and minimize the harm from those threats.[8] The use of machine learning raises the possibility of automated self-defense systems in a much wider variety of contexts, systems that could draw on the resources of the entire armed forces and possess the capacity to launch counter-attacks. This is a new challenge that scholars have not yet addressed.

Second, the increased use of machine learning-produced analysis in the (human) decision-making process raises questions about the quality and reliability of the analysis, and the level of deference that humans will give to machine learning-produced recommendations (i.e., whether a human will be willing to overrule the machine). These issues are not new in

[7] Cite. [Perhaps use the role of intelligence briefings in drone strike decisions.]

[8] *See, e.g.*, Paul Scharre, *Presentation at the United Nations Convention on Certain Conventional Weapons*, p. 3, Apr. 13, 2015, https://www.unog.ch/80256EDD006B8954/(httpAssets)/98B8F054634E0C7EC1257E2F005759B0/$file/Scharre+presentation+text.pdf.

and of themselves – questions about the reliability of intelligence sources and the weight to give to a particular recommendation have long existed – but, as we show herein, machine learning techniques raise new questions and add layers of complexity to existing debates.

Third, discussions regarding machine learning inevitably give rise to questions about the explainability and transparency of the decision-making process. On the one hand, the nature of machine learning means that the reasons underpinning a machine's particular recommendation or exercise of automated decision-making may not be transparent or explainable. While a lack of public-facing transparency regarding use of force decisions is not new, at issue here is the quality and interrogability of the recommendation that the machine produces for the decision maker. The state may therefore not only be unwilling to explain the reasons behind a particular course of action it takes, it may be unable to do so. On the other hand, if a state can program an algorithm to explain its recommendations, this may actually increase a state's ability to explain the rationale underpinning its decision to use force. In this context, machine learning again adds complexity to existing debates.

This essay's goal is to draw attention to current and near future developments that may have profound implications for international law, and to present a blueprint for the necessary analysis. More specifically, this essay seeks to identify the most likely ways in which states will begin to employ machine learning algorithms to guide their decisions about when and how to use force, to identify legal challenges raised by use of force-related algorithms, and to recommend prophylactic measures for states as they begin to employ these tools.

Part II sets out specific scenarios in which it seems likely that states will employ machine learning tools in the use of force context. Part III then explores the challenges that states are likely to confront when developing and deploying machine learning algorithms in this context, and identifies some initial ways in which states could address these challenges.

In use of force decisions, the divide between policy and legal doctrine is often disputed. Some may consider aspects of this essay to be matters of policy, while others may consider those same aspects to be matters of legal obligation. The purpose of the essay is not to resolve those debates, but rather to focus on emerging concepts and challenges that states should consider whenever artificial intelligence plays an influential role in their decisions to use force, in light of the international legal framework.

II. RESORT TO FORCE SCENARIOS

States must make a variety of calculations when confronted with a decision about whether to use force against or inside another state. The UN Charter provides that states may lawfully use force in self-defense if an armed attack occurs.[9] This means that a state that fears that it is likely to be the subject of an armed attack will urgently want to understand how likely it is that the threatened attack will manifest in actual violence. Equally, a state subject to an armed attack will seek to understand which actor attacked it, whether it appears necessary to use force in response to that attack, and whether its response would be proportionate.[10] In situations such as these, it appears increasingly likely that states will seek to use machine learning technology to strengthen their decision-making processes, both in a traditional reactive sense and in a predictive context. War-gaming exercises provide a straightforward example that demonstrates the potential utility of machine learning.[11] These exercises currently rely on a form of role-playing of adversaries, and could further be

[9] UN Charter Art. 51. Under customary international law, acts of self-defense must be both necessary and proportional. Military and Paramilitary Activities in and Against Nicaragua, Nicaragua v. United States, Merits, [1986] ICJ Rep. 14, para. 176.

[10] NOAM LUBELL, EXTRATERRITORIAL USE OF FORCE AGAINST NON-STATE ACTORS (2010), chapter 2.

[11] Cite.

enhanced and transformed by machine learning systems that are able to provide modelling and simulations based on advanced analysis of capabilities and behaviour patterns, and thus provide a more accurate tool for assessing the flow of events.

Beyond these general use of force inquiries, there are two other contexts in which machine learning might come into play. The first context is in cyber operations: states may begin to produce and rely on machine learning-driven algorithms that allow them to defend against cyber attacks at the speed of light, in what may come to look like automatic self-defense. The second context is somewhat novel, but is likely to represent an emerging threat. It is possible that states and nonstate actors may use machine learning to forge messages—such as video commands ordering an attack or retreat, statements by politicians, or threats of war—in order to gain an advantage in the context of both the resort to force and the conduct of armed conflict.[12] By employing sophisticated video and audio forgeries enhanced by recent advances in machine learning, an actor may attempt to manipulate the target state's chain of command or public opinion. This possibility raises a number of new challenges, and brings into play several rules of international law, but is distinct from the self-defense framework on which this essay focuses, so we do not discuss it further herein.

A. Informing the Exercise of Self-Defense

States may use machine learning to help them decide when to act in self-defense. There is a longstanding debate about whether and when it is permissible to use force before

[12] *See* Sven Charleer, *Family fun with deepfakes. Or how I got my wife on the tonight show*, MEDIUM, Feb. 2, 2018; Jon Christian, *Experts fear face swapping tech could start an international showdown*, THE OUTLINE, Feb. 1, 2018.

an armed attack is actually completed.[13] A variety of states and scholars accept that a state may act in situations in which the attack has not yet occurred,[14] but where the need to respond is "instant, overwhelming, and leaving no choice of means, and no moment for deliberation."[15] Others go further, arguing that in some contexts it may be appropriate for a state to act to prevent a particular, tangible, and serious threat from turning into an armed attack, even if the attack is not imminent.[16] Some have referred to the permissibility of acting in the "last feasible window of opportunity" before the attack occurs, particularly when the threat is posed by cyber operations or weapons of mass destruction.[17]

As a result, it is critical for a state to reasonably assess when an armed attack is imminent. Machine learning may help a state assess the statistical likelihood that certain pre-attack activities by another state will develop into an actual attack, and help a state assess the

[13] *Compare* DEREK BOWETT, SELF-DEFENCE IN INTERNATIONAL LAW 191-92 (1958), *with* IAN BROWNLIE, INTERNATIONAL LAW AND THE USE OF FORCE BY STATES 275-78 (1963); Sean Murphy, *The Doctrine of Preemptive Self-Defense*, 50 VILL. L. REV. 699, 703 (2005).

[14] Christopher Greenwood, *International Law and the Pre-Emptive Use of Force: Afghanistan, al Qaida, and Iraq*, 4 SAN DIEGO J. INT'L L. 14-15 (2003) (listing Franck, Waldock, Fitzmaurice, Bowett, and others as supporting anticipatory self-defense).

[15] Letter from Daniel Webster, U.S. Secretary of State, to Lord Ashburton, British Plenipotentiary (Aug. 6, 1842), *quoted in* II JOHN BASSETT MOORE, A DIGEST OF INTERNATIONAL LAW (1906), § 217, at 412.

[16] *See* Ashley Deeks, *Taming the Doctrine of Pre-Emption*, *in* THE OXFORD HANDBOOK OF THE USE OF FORCE IN INTERNATIONAL LAW 666-67 (Marc Weller ed., 2015) (discussing literature and state practice supportive of pre-emptive self-defense); Noam Lubell, *The Problem of Imminence in an Uncertain World*, *in* THE OXFORD HANDBOOK OF THE USE OF FORCE IN INTERNATIONAL LAW 697, 702 (Marc Weller ed., 2015) (discussing arguments for and against expanding the traditional concept of imminence).

[17] *Cf.* MICHAEL SCHMITT, ED., TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS, Rule 73 at 351 (citing "last feasible window of opportunity" favorably in the cyber context).

gravity of a particular, looming threat that might materialize.[18] Additionally, a state might use machine learning algorithms to assist it in performing risk calculations about the levels of force that might be used by it and against it. For instance, a state that fears an impending armed attack might seek to calculate the risk of false positives and false negatives—that is, how much damage it might suffer if it does not use force anticipatorily, how much damage it might cause if it acts in anticipatory self-defense unnecessarily, and how much damage its enemy likely would inflict on it if it were to act in anticipatory self-defense.[19]

Risk analysis has long been recognised as a crucial tool in decision-making in other spheres of government, from health care provision to the criminal justice system,[20] but it appears to be comparatively under-utilized in government decisions on the use of force. Notwithstanding the many academic writings that exist on risk analysis in going to war, it is questionable whether high-ranking political and military officials today undertake the kind of detailed risk reviews that happen in other fields. In health care, for example, an analysis of the statistical probabilities and cumulative effects of false positives and false negatives are

[18] It is possible to draw a potential parallel here with recent developments in the game of Go. Advances in artificial intelligence systems have allowed scientists to produce software that can beat the game's human world champion. The system also can improve its own capabilities by playing against itself. The system runs through millions of patterns and variations and selects options that are mostly likely to result in particular outcomes (in this case, victories). James Vincent, *DeepMind's Go-playing AI doesn't need human help to beat us anymore,* THE VERGE, Oct. 18, 2017. A machine learning system employed in anticipatory self-defense contexts could run through millions of possible types of attacks and possible responses and attach probabilities to each outcome.

[19] For an analysis that combines psychology and policy decisions in this context, see Philip Dunwoody & Kenneth Hammond, *The Policy of Preemption and Its Consequences: Iraq and Beyond*, 12 J. PEACE PSYCH. 19 (2006). For a general discussion of false positives and negatives, see KENNETH HAMMOND, HUMAN JUDGMENT AND SOCIAL POLICY: IRREDUCIBLE UNCERTAINTY, INEVITABLE ERROR, UNAVOIDABLE INJUSTICE (1996), ch 1.

[20] *See* HAMMOND, HUMAN JUDGMENT, *supra* note 19 (discussing uncertainty in decision-making).

commonplace in policy-making.[21] The proliferation of machine learning might prompt political and military officials to follow course and rely more heavily on detailed machine learning risk assessments when deciding whether and how to use force.

One reason to think that states may employ machine learning this way is that states are already seeking to create "early warning systems," of which pre-attack detection might be one type.[22] As one former intelligence analyst noted, "Intelligence agencies use advanced algorithms to interpret the meaning of intercepted communications, identify persons of interest, and anticipate major events."[23] Machine learning algorithms crafted to anticipate the fact and location of particular threatening geopolitical developments—such as the movement of missile launchers or increased chatter over terrorist networks—might also share some commonalities with existing risk assessment algorithms that national and local governments use to allocate law enforcement resources toward particular locations where crime is most likely to occur.[24] Indeed, in one recent political science study, scholars were able to use machine learning tools to identify changes in news coverage in authoritarian regimes in the lead-up to armed conflict.[25] Tools like these might provide critical information to states that

[21] MICHAEL DRUMMOND ET AL., METHODS FOR THE ECONOMIC EVALUATION OF HEALTH CARE PROGRAMMES (4th ed. 2015).

[22] Defense Science Board, Summer Study on Autonomy 2 (June 2016) (discussing "[e]arly warning system for understanding global social movements"); *id.* at 4 ("Imagine if we had autonomous high performance computing engines capable of not only searching 'big data' for indicators of WMD proliferation, but of deciding what databases to search... to provide early warning and enable action.").

[23] Cortney Weinbaum, *The Ethics of Artificial Intelligence in Intelligence Agencies*, NATIONAL INTEREST, July 18, 2016.

[24] *See, e.g.,* Justin Jouvenal, *Police Are Using Software to Predict Crime. Is It a Holy Grail or Biased Against Minorities?*, WASH. POST, Nov. 17, 2016.

[25] Elena Labzina & Mark D. Nieman, Understanding the Use of State-Controlled Media: A Machine-learning Approach, Jan. 2017.

believe that they are facing an imminent armed attack. Of course, a number of machine learning-based risk assessments have faced criticisms regarding the quality and reliability of outputs, as well as potential bias. States should address such concerns before they deploy these systems in a use of force context, an issue we discuss further below.[26]

Whether there is sufficient (and sufficiently accurate) historical data around which one could build predictive algorithms today is unclear. One significant challenge is that historical data is harder to locate and, because it is often not in an electronic format, harder to use than contemporary data. (Unlike twenty years ago, most actors today keep electronic records of their activities.) However, a U.S. Department of Defense report recently argued that the capability to develop early warning systems

> may soon be achievable. Massive datasets are increasingly abundant and could contain predictive clues . . . . Recent advances in data science, ever-improving computational resources, and new insights into social dynamics offer the possibility that we could leverage these massive data sources to make actionable predictions about future world events.[27]

The United Nations, the U.S. Central Intelligence Agency, and the U.S. Department of Defense are already attempting to parse public data to anticipate when and where conflict may arise.[28] As governments continue to collect vast quantities of data on a wide variety of topics, it is increasingly likely that they will begin to train predictive algorithms in support of use of force decision-making in the near future.

[26] *See infra* Part III.

[27] Defense Science Board, *supra* note 22, at 79.

[28] Sheldon Himelfarb, *Can Big Data Stop Wars Before They Happen?*, FOREIGN POLICY, Apr. 25, 2014.

B. Automaticity in Self-Defense

The examples described in section A represent cases in which machine learning algorithms are used to *inform* a state's response to a particular scenario. However, particularly in the cyber context, states are likely to enable machine learning-driven algorithms to *respond themselves* to incoming cyber operations that take place at the speed of light. That is, states may grant algorithms a direct role in decision-making. To the best of our knowledge, most current cyber defenses engage primarily in self-preservation against incoming attacks, for example by erecting firewalls and shutting themselves off from outside networks to prevent further damage. Notwithstanding the current reactive posture, it is entirely possible that in certain circumstances states could program their systems to respond with a form of counter-attack, a "hack-back" of sorts, to disable the systems launching the cyber-attack. These cyber operations implicate the use of force to the extent that the offensive cyber operations constitute (cyber) armed attacks and the responsive cyber operations represent acts of self-defense. Of course, it is important to distinguish among different types of responses to cyber armed attacks, and to note that not all responses will qualify as uses of force necessitating a self-defense-based justification.[29]

This use of machine learning seems relatively near at hand, based on publicly stated concerns by military officials in cyber-capable states. For instance, an Obama-era senior acquisitions official at the U.S. Department of Defense argued that "making our robots wait for human permission would slow them down so much that enemy AI without such

[29] Although this article does not focus on responses to international law violations that fall below the level of the use of force, it is quite possible that states also will choose to employ machine learning systems to help them calculate what types of lawful counter-measures would be most effective in response to such violations.

constraints would beat us."[30] A Pentagon Defense Science Board study discusses "automated cyber responses."[31] Significantly, the U.S. Department of Defense Directive on Autonomy in Weapons Systems does not apply to autonomous cyberspace platforms for cyber operations, leaving the U.S. military less constrained in developing and employing cyber weapons that might act in automated self-defense.[32]

Creating a lawful system of automated cyber self-defense that can respond in a timely manner to incoming attacks will require the algorithm's programmers to (a) determine whether the incoming cyber operations amount to an armed attack, a use of force, or another form of unlawful intervention; (b) initiate a corresponding automated cyber response that qualifies as an act of self-defense or counter-measure, depending on the nature of the incoming operation;[33] and (c) ensure that the nature of the response adheres to the applicable legal rules, such as necessity and proportionality for self-defense.[34] Given the speed at which cyber operations occur, automation may be necessary: human involvement may be simply impossible, at least in certain respects. However, if two states both employ algorithms that act

[30] Sydney Freedberg, Jr., *Should US Unleash War Robots? Frank Kendall vs. Bob Work, Army*, BREAKING DEFENSE, Aug. 16, 2016.

[31] Defense Science Board, *supra* note 22, at 2; *id.* at 4 ("Imagine if . . . [w]e had an autonomous system to control rapid-fire exchange of cyber weapons and defenses, including the real-time discovery and exploitation of never-seen-before zero day exploits...enabling us to operate inside the 'turning radius' of our adversaries.").

[32] Department of Defense Directive 3000.09 (Nov. 21, 2012), section 2.b; *see also* DEFENSE ONE, *Trump's Pick for NSA/CyberCom Chief Wants to Enlist AI For Cyber Offense*, Jan. 9, 2018. The UK National Cyber Security Strategy 2016-2021 indicates a UK interest in autonomous cyber systems. It states that the government will continue to fund further cyber research, including on "big data analytics [and] autonomous systems." Para. 7.3.6.

[33] *See supra* note 29.

[34] The algorithm likely also would need to be programmed to respond in a manner consistent with the laws of armed conflict, including distinction and proportionality.

in automated self-defense, they may find themselves in an armed conflict without either immediately knowing it. It presumably will be possible for machine learning algorithms to game out various responses (to and fro between the parties) and to both select the best one for the attacked state and to identify potential consequences (facilitating some form of early warning, or human-focused alert). Such systems would not be dissimilar to the algorithms that are performing so successfully at Go.[35] It presumably also will be possible for programmers to establish certain limits to the algorithm's responses, although—as discussed in the next part—there are situations in which algorithms act unpredictably when confronting other algorithms.[36] Given the risk of escalation, it seems essential that safeguards be put in place, which may include limiting the parameters within which an algorithm can act or triggering an alarm when certain conditions are present.

Although cyber operations are the most obvious situations in which states may choose to employ algorithms that enable automatic responses (some of which may require claiming a legal right to self-defense), similar issues might arise in the context of automated air defense systems or in confrontations between unmanned aerial or underwater vehicles. Using machine learning tools, states may construct these systems to allow an unmanned underwater vehicle, say, to identify the object it is confronting, make sense of the object's movements, determine whether the object is a foreign, unmanned weapons platform, and respond almost instantaneously to a missile fired from that foreign platform.[37] It may also be possible for programmers to incorporate the concept of an imminent threat, which would enter the realm

[35] ECONOMIST, *The Latest AI Can Work Things Out Without Being Taught*, Oct. 21, 2017.

[36] *See* Mark Buchanan, *Physics in Finance: Trading at the Speed of Light*, NATURE, Feb. 11, 2015 (discussing how algorithms can interact with each other in unforeseen ways).

[37] For a discussion of an existing Chinese underwater unmanned vehicle, see Defense Science Board, *supra* note 22, at 43. These systems operate in complicated dynamic environments, and as such are distinguishable from static automatic self-defense systems such as the Phalanx Close-In Weapons System.

of automatic anticipatory self-defense, or to link up different systems, facilitating automated counter-offensive operations.[38] Although it would be possible to keep a "human on the loop" in both the cyber and unmanned vehicle contexts, "the value of such algorithms resides in their ability to initiate immediate responses. Once an AI system detects an imminent attack, its value would reside in the ability to respond before the attack could take place."[39] The temptation to rely on the algorithm alone to guide decision-making therefore will be powerful.

C. Informing Necessity and Proportionality Analyses

Before using force in self-defense, a state must determine that the use of force is necessary (that is, that no other peaceable options exist and that the force is limited to that which is required).[40] Assuming that force is necessary, that state also must limit its response to that which is proportionate to defeat the attack or the threat raised thereby.[41] As with predictions about whether and when an attack might occur, states may turn to machine learning to help them predict with greater certainty whether force is necessary and what levels of responsive force would be proportionate.

It is precisely this ability to work through a range of different options and likely associated outcomes that represents a key added value of machine learning. For instance,

[38] For instance, a state's response may extend to target selection, through incorporation of the necessity and proportionality analyses discussed in Part III.C.

[39] Weinbaum, *supra* note 23.

[40] *Advisory Opinion on the Legality of the Threat or Use of Nuclear Weapons*, [1996] I.C.J. Rep., para. 41; JUDITH GARDAM, NECESSITY, PROPORTIONALITY, AND THE USE OF FORCE BY STATES 4 (2004); International Law Association, *Report on Aggression and the Use of Force* (forthcoming 2018) (draft on file with authors).

[41] BROWNLIE, *supra* note 13, at 261.

algorithms might help a state calculate whether means other than force might be sufficient to address an attack, or algorithms may recommend which targets, if hit, might most effectively suppress the forcible conduct. For example, the algorithm might recommend that the defending state target a particular command center rather than a particular missile installation to most effectively stop the attack. The use of machine learning in self-defense is likely to form part of an interconnected system: one algorithm might indicate that a forcible response is necessary, while a different algorithm might help the state calculate the predicted effects of its response, thereby facilitating the state's compliance with the proportionality rule.

D. Masking the Source of an Attack to Aid Aggression

Cyber conflict presents a number of new challenges that states must address.[42] One possibility, particularly relevant to the use of machine learning, is that an attacker may launch certain attacks in order to provoke a (potentially automated) self-defense response by the victim state.[43] For instance, an attacker may conceal its own act of aggression by taking over systems in a third state and using these systems to launch attacks on the victim state. This is a relatively straightforward case of deception. However, assume an attacker knows that the victim state employs defensive machine learning algorithms. The attacker may attempt to further manipulate the situation by not just seeking to mask its own use of force, but by attempting to trigger an act of self-defense by the victim state against the third state that will necessitate a response by the third state, resulting in an armed conflict between the victim and third states. This strategy may attempt to manipulate (or "game") the victim's machine

[42] For initial efforts in this regard, see TALLINN 2.0, *supra* note 17.

[43] Phil Stewart, *Deep in the Pentagon, a Secret AI Program to Find Hidden Nuclear Missiles*, REUTERS, June 5, 2018.

learning algorithm, to increase the likelihood of the desired response. For instance, the attacker could utilise critical infrastructure in the third state to launch its attack, so that the response by the victim state would cause harm to the third state's critical infrastructure. This, in turn, increases the likelihood that the third state will be pushed into responding to any use of force by the victim state.

The attacker may seek to further increase the likelihood of this outcome by incorporating a combination of other means, such as fake communications chatter in the victim and third states (indicating increased planning activity, etc.), the use of video and audio forgeries, and the large-scale deployment of bots to contribute to increased public discourse.[44] As all of these sources are likely to provide input data for a self-defense algorithm, they may contribute to the manipulation of any resultant risk scores. False-flag deception by militaries has long existed,[45] but the use of machine learning makes such tactics both more likely and more likely to be effective, while raising concerns regarding the potential for "gaming" automated response systems.

III. CHALLENGES SURROUNDING MACHINE LEARNING AND THE LAW ON THE RESORT TO FORCE

The previous part discussed various scenarios in which states may seek to incorporate machine learning into their use of force decisions. While it is likely that states will

[44] For a discussion of the use of internet bots to spread misinformation, see *First Evidence that Social Bots Play a Major Role in Spreading Fake News*, MIT TECH. REV., Aug. 7, 2017, https://www.technologyreview.com/s/608561/first-evidence-that-social-bots-play-a-major-role-in-spreading-fake-news/.

[45] *See* YORAM DINSTEIN, THE CONDUCT OF HOSTILITIES UNDER THE INTERNATIONAL LAW OF ARMED CONFLICT 274 (3d ed. 2016).

increasingly deploy machine learning technologies in this context, their use poses technological, legal, and ethical challenges. In light of the types of machine learning that states might employ to guide their use of force decision-making, this Part identifies the most salient challenges states will confront and offers suggestions for how states might address these concerns.

A. Algorithmic Quality and Reliability

A number of factors are relevant to the quality and reliability of an algorithmic process, including: whether it is actually possible for an algorithm to perform the desired task, whether there is sufficient quality data to allow the algorithm to perform its task, and the extent to which programmers can reproduce legal rules and judgment in code. These factors should in turn influence both a state's decision to deploy an algorithm and the manner in which the state deploys it.

1. Is it possible for an algorithm to perform the desired task?

Before developing or deploying an algorithm in a use of force context, the first question to ask is whether it is actually possible, in principle, for an algorithm to perform the required task. The answer depends upon both the nature of machine learning and the nature of the situation to which machine learning is applied.

Machine learning operates on the basis of correlation, not causation.[46] As such, the algorithms analyse input data points against a specified output to identify patterns of correlation. Put simply, this enables the algorithm to establish that, if certain input factors are

[46] *See, e.g.*, Anjanette H. Raymond et al., *Building a Better HAL 9000: Algorithms, The Market, and the Need to Prevent the Engraining of Bias*, 15 NW. J. TECH. & INTELL. PROP. 215, 224-25 (2018).

present, then it is likely—to a certain degree of accuracy—that a particular output or consequence will follow. The accuracy of the output is affected by both the quality and quantity of the input data.[47] Certain problems are more amenable to this type of reasoning than others. A (simplified) distinction may be made between problems that involve logical patterns, or some form of cause and effect, and those that involve inherent uncertainty or randomness.[48]

Applied to the use of force, it is easy to see that machine learning algorithms can be deployed straightforwardly to detect live cyber-attacks, as these would be relatively measurable and more obviously likely to cause harm. Such attacks involve communications that flow across the architecture of the Internet, and that may be detected by identifying a particular pattern or profile associated with a specific attack, or by noting unusual communications patterns.[49] Out of the ordinary activity may indicate that an attack is either in preparation or actually occurring. Conversely, algorithms such as those that make recommendations regarding whether a use of force is proportionate may face greater challenges. These predictions appear closer to the kinds of modelling that militaries currently conduct when estimating collateral damage. However, proportionality in the use of force context entails more than just a calculation of anticipated harm from specific military operations, and must include wider assessment of overall harm to the affected state relative to

[47] Ziad Obermeyer & Ezekiel J. Emanuel, *Predicting the Future: Big Data, Machine Learning, and Clinical Medicine*, 375 N. ENGL. J. MED. 1216 (2016).

[48] This arises due to the fact that, as noted above, algorithms typically work on the basis of correlation and not causation.

[49] Public-Private Analytic Exchange Program, *Cyber Attribution Using Unclassified Data*, p. 17 (2016), https://www.dni.gov/files/PE/Documents/Cyber-Attribution.pdf.

aims of the defensive response.[50] An algorithm would therefore need to be far more sophisticated than current military models. That said, although the problem is more complex than making collateral damage estimates, it is of a similar nature, and an effective solution may only be a matter of time.

The use of machine learning to predict whether a state is likely to launch a traditional kinetic attack is equally problematic. The circumstances in which states use force are so vastly diverse that comparisons across examples may be irrelevant. Equally, a significant element of uncertainty and unpredictability often surrounds the decision to use force. For instance, one state may undertake aggressive posturing or brinksmanship to indicate that it is *prepared* to use force or to prompt a particular reaction in a rival state, even though it is not actually planning to undertake forcible acts.[51] Internal or international political dynamics—and even inherent human unpredictability—may also affect any decision to use or threaten force. As such, any algorithm analysing this situation must not only take into account obvious matters such as troop movements or weapons stockpiling, but also factors such as political dynamics between and within states, economic interests, and human psychology. The use of an algorithm to predict the likelihood that a state will use force is therefore far from straightforward.

2. Is there sufficient high-quality data for the algorithm to perform its task?

The next question is whether there is data of sufficient quality to allow the algorithm to perform its task to the required degree of accuracy and reliability. One key issue is the availability of training data. As noted, machine learning algorithms work on the basis of

[50] Amichai Cohen & Noam Lubell, *Strategic Proportionality as a Regulatory Mechanism for Armed Conflict* (forthcoming 2018).

[51] Iraq's troop movements towards the Kuwaiti border during the 1990s offer one such example.

correlation and so sufficient training data (that is, input data linked to associated outcomes) is required in order to train the model so that it can be applied to new situations.[52] Again, while this may not be a problem in the context of cyber operations, it is likely to be a significant factor with respect to the decision to launch an armed attack. Although there have been far too many casualties of war in recent decades, the number of incidents internationally recognized as qualifying as armed attacks since the UN Charter came into force is relatively small.[53] This will affect the accuracy of any algorithmic decision, and may indicate that states should postpone creation and deployment of this type of algorithm until sufficient data is available.

Embedded in the question of reliability is the challenge of determining precisely how accurate algorithms should be before states feel comfortable using them. This question has arisen in the autonomous weapons context as well.[54] Must the algorithm produce results that are as good as those humans produce before a state uses it? Should a state only employ algorithms that make *more* accurate predictions than humans make? States will need to answer these questions at the same time that they assess the quality and quantity of data needed to produce useable algorithms, and the degree of explainability required; that is, they need to define what a "useable" algorithm is.

An added complexity with respect to algorithmic quality is the issue of bias.[55] This relates to both input data and the operation of the algorithm itself. For instance, the state (and indeed the programmer herself) will inevitably incorporate institutional values and

[52] Add cite.

[53] For a relatively complete list of armed attacks since 1945, see TOM RUYS & OLIVIER CORTEN EDS., THE USE OF FORCE IN INTERNATIONAL LAW: A CASE-BASED APPROACH (2018).

[54] Michael Horowitz & Paul Scharre, *The Morality of Robotic War*, N.Y. TIMES, May 26, 2015.

[55] Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972 (2017); Matthias Spielkamp, *Inspecting Algorithms for Bias*, MIT TECH. REV., June 12, 2017.

preferences into code. A self-defense algorithm of a state that is deeply hesitant to use force, or that analyzes situations through a particular world view, will operate differently than an algorithm of a state that is more willing to employ force (and interprets the Charter and customary law as more permissive), or that analyses situations through a different political lens. These two states would need (and elect) to train their algorithms on different kinds of historical examples that reflect their own preferred approaches to self-defense. In addition, threat assessments will, as stated in the previous section, require an analysis of the political discourse and cultural, social, and psychological profiling of decision-making dynamics in other states. This creates an additional risk of programming bias in relation to the way algorithms are designed to interpret actions of specific states, because states may incorporate (possibly faulty) preconceptions about other states into the machine learning process.

3. The difficulty of encoding legal rules

Another significant challenge—one faced by anyone creating an algorithm that makes recommendations driven by underlying bodies of law—is the difficulty of translating broad legal rules into very precise code.[56] In the use of force context, an algorithm created to help a state respond in self-defense would need to reflect the appropriate international law triggers: a determination that an armed attack has occurred or is imminent.[57] A programmer crafting an algorithm that a state might use to help it gauge the lawful scope of a forcible defensive response would need to convert contested and malleable concepts such as necessity and

[56] Maayan Perel & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STAN. TECH. L. REV. 473, 518 (2016 ("This process of translating legal mandates into code inevitably embodies particular choices as to how the law is interpreted, which may be affected by a variety of extrajudicial considerations, including the conscious and unconscious professional assumptions of program developers.").

[57] UN Charter Art. 51; Murphy, *supra* note 13.

proportionality into discrete, binary decision trees. This already has proven a challenge in domestic legal systems that employ algorithms to make decisions about how to allocate welfare benefits and who to exclude from air travel, for instance.[58] It is likely to prove an even harder problem in international law, because concepts often exist at a higher level of generality and because there may be less consensus among stakeholders about how to interpret these concepts. The approach that a state takes to international law, including in the realm of use of force, rarely takes the form of a rigid set of rules, is usually context dependent, and often is subject to a range of views and changing administrations. Efforts to transform this into code would therefore require constant debate and the ability to continuously edit and change fundamental sections of the algorithm. Further, coders are unlikely to have experience with or training in law, making the translation exercise fraught.

4. What does this mean for the use of algorithms?

The above discussion indicates that states should take a number of factors into account when considering the use of machine learning algorithms in a use of force context. An initial question is whether the algorithm can actually perform the desired task. There is a distinction between the use of algorithms to *make* a decision and the use of algorithms to *inform* a decision. The latter is far less radical than the former. A follow-on step is then to ensure that the algorithm has sufficient quality of data, including training data, and is coded appropriately, so that the state can ensure its quality and reliability, or at least quantify those characteristics. The state additionally must ensure that the algorithm's users know the circumstances in which the algorithm can work, its level of accuracy, and what it can or cannot do. This will help to ensure that those deploying algorithms, or making decisions

[58] Danielle Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1252 (2008).

informed by algorithms, are aware of the algorithms' limitations and can effectively evaluate outputs.[59]

States might take a number of other measures to enhance quality and reliability. For instance, states might hard-code the prioritisation of particular (reliable) intelligence sources into the design; might identify acceptable levels of accuracy, and therefore of false positives and false negatives; and might facilitate close interaction between developers and international lawyers during the design process.

B. Level of Human Involvement

Whenever government officials deploy algorithms to guide their decision-making, they must determine how heavily to rely on those algorithms in weighing courses of action. For several of the use of force scenarios described in Part II, machine learning tools seem likely to provide valuable guidance, but are unlikely to receive absolute deference from states' leadership and their militaries. That is, states may use algorithms to inform and supplement human decision-making, but generally are unlikely to grant those algorithms decision-making authority. The "automatic self-defense" cyber scenarios discussed above, however, pose a deeper challenge, one that potentially pits practicality against ethics: Should war-and-peace decisions necessarily be made by humans?

1. The concept of "meaningful human control"

[59] For further discussion regarding the use of algorithms, see Lorna McGregor, Daragh Murray & Vivian Ng, *International Human Rights Law as a Framework for Algorithmic Accountability* (forthcoming 2018).

States, scholars, and ethicists have begun to consider these questions in the heated debate about lethal autonomous weapons systems, sometimes referred to pejoratively as "killer robots." As some anticipate a future in which states will deploy fully autonomous robots to identify and kill combatants, various actors have argued that it is imperative for a human to maintain "meaningful human control" over decisions to employ force during armed conflict.[60] The use of algorithms in relation to the initial decisions on resort to force raise similar questions, perhaps even with added urgency. If the possibility that a machine might be given the power to "decide" to kill a single enemy soldier is fraught with ethical and legal debates, what are we to make of the possibility that a machine could ultimately determine whether a nation goes to war, and thus impact thousands or millions of lives?

Disappointingly, a closer look at "meaningful human control" in the autonomous weapons discourse reveals it to be deceptively alluring but substantively vague and unhelpful. A number of actors and scholars have attempted to define the concept, focusing upon elements such as the nature of the role the human plays in the use of such weapons; the availability of appropriate information for the human operator; and the effective control the human has over the system.[61] All these elements are, however, subject to differing interpretations that become even more apparent when one moves from the debate over autonomous weapon systems into the use of algorithms in the *jus ad bellum*.

In the autonomous weapons sphere, one issue at stake is whether a machine should be allowed to have the final "say" in whether to pull the trigger and kill an individual.

[60] *See, e.g.*, UNIDIR, The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward (2014).

[61] *See, e.g.*, Michael C. Horowitz & Paul Scharre, *Meaningful Human Control in Weapon Systems: A Primer*, Center for a New American Security, Working Paper, March 2015; *see also* Human Rights Watch, *Killer Robots and the Concept of Meaningful Human Control, Memorandum to Convention on Conventional Weapons (CCW) Delegates*, April 2016.

Reasonable arguments have been made pointing out advantages and disadvantages to the various approaches,[62] but ultimately this appears to be a question of policy and ethics rather than law. While it remains an open ethical question in the autonomous weapons sphere, the corresponding issue is less likely to be a debate for the *jus ad bellum*. It may be theoretically correct that in certain circumstances we would be safer by placing our trust in a well-designed algorithm than in allowing irrational and unpredictable leaders to control nuclear red buttons. Nonetheless, a world in which the final decision on taking the nation to war is delegated to a computer is hard to imagine, and while the future may hold surprises, for the time being such scenarios can be filed under the fictional part of science fiction. At least two situations do emerge, however, that pose challenges vis-à-vis human involvement. The first is the use of algorithms to inform the decision-making process. The second is the potential need to allow algorithms to make decisions (that may rise to the level of the use of force) in particular contexts, notably cyber.[63]

2. Algorithms that inform

Algorithms could—and arguably are likely to—play a crucial role in almost every part of the decision-making process leading up to the point when a state initiates force. Their use is likely to grow in particular in the area of threat assessment, thereby providing the foundations upon which humans make the decision to go war. Machine calculations could consequently play a critical role in life and death decisions for whole countries, impacting far more lives than the autonomous weapons currently being so vigorously debated. This is perhaps a natural result of the tendency to pay more attention to individualised contexts than

[62] *Id.*

[63] *See supra* Part II.B.

to abstract policy. For example, introducing algorithms that calculate survival chances and cost-effectiveness to determine whether to pull the plug on a single terminally ill individual would probably horrify many, even though we might accept using algorithms for cost-effectiveness calculations that affect the provision of healthcare and medicines to millions.

In the context of resort to force, the role of algorithms in the underlying calculations could lead states to unwittingly make war-related decisions almost entirely based on machine calculations and recommendations.[64] A key concern is that, even if the final war-making decision rests with humans (in this case, a state's executive branch), the degree to which that human involvement is considered "meaningful" may be questioned. Two relevant factors are (a) the centrality of algorithms to the decision-making process, and (b) the deference given to algorithmically-produced decisions.

First, as algorithms become increasingly central to state decision-making processes, the number of algorithmic recommendations that inform—and therefore influence—a final decision is likely to increase. This means that the impact of any inadequacy within the underlying algorithm[65] may be magnified. The potential exists for a cascade effect, whereby a small bias in the first algorithmic recommendation is fed into a second algorithmic recommendation, which in turn skews the input for a third recommendation, and so on. Within any inter-connected and inter-dependent process, the impact of even a small error may be extensive.[66] This means that although a human may make the final decision, the basis on

[64] In these scenarios, the state presumably has decided not to deploy an algorithm in a decision-making role. For the purpose of this example, questions of reliability etc., although pertinent, are secondary to the question of whether the level of human involvement is meaningful.

[65] *See supra* Part III.A.

[66] Any error will be compounded should bias or other inadequacy exist with respect to each algorithmically informed recommendation in the decision-making chain.

which that decision is made, and the range of available options, may have been heavily influenced by the underlying algorithm(s).

The other factor is the deference granted to algorithmically-produced decisions, a phenomenon often referred to as "automation bias." Some studies have found that operators of automated systems tend to trust those systems' answers,[67] including in circumstances in which their own judgment or experience would have led to a different conclusion.[68] One reason for this may be that people presume that algorithmic decisions are made on the basis of indisputable hard science, or operate at a level beyond human capacity, or because people fear overruling the computer and "getting it wrong." This phenomenon is likely to be exacerbated in the context of use of force decisions. It is in the very nature of algorithmic systems to engage in complex and potentially undecipherable calculations, and to compute across a scale of factors and at a speed that humans cannot replicate.[69] Even if one takes the position that the decision to use force at the state level must not be delegated to machines, it is hard to escape the fact that algorithms are likely to play a growing role in all of the underlying calculations. The more advanced algorithms become, the less meaningful human decisions become. Ensuring the algorithms' quality and reliability, as well as transparency and explainability, will therefore be of utmost importance.[70] States must also take the human factor into consideration. The likelihood that a human operator will defer to an algorithmic decision may only increase in the resort to force context, as the potential consequences associated with overruling an algorithmic recommendation and deciding *not* to use force are

[67] Citron, *supra* note 58, at 1271 and notes 147-48.

[68] Lonnie Shekhtman, *Why do people trust robot rescuers more than humans?,* CHRISTIAN SCI. MONITOR, Mar. 1, 2016.

[69] The inability to understand the calculations is likely to intensify the more states move toward machine learning systems that can create new algorithms beyond the original human input.

[70] *See infra* Part III.C.

significant, to say the least. As a result, individuals should be trained to resist unwarranted automation bias.

3. Algorithms that make direct decisions

In certain contexts, transferring decision-making to an algorithm may be necessary. This is most obviously the case in the cyber sphere, though it is possible that the need may arise in other cases as well, such as in situations of leadership decapitation. When computer systems come under a cyber-attack, the most efficient response will often be an immediate one generated by the machine itself. Indeed, operations may occur at such speed that human decision-making is rendered inadequate. In most cases this will involve a system undertaking purely defensive measures to prevent further intrusion and damage, such as closing itself off from the network.[71] However, it is also possible that a state could craft an automated cyber counter-attack to halt an ongoing attack, or to harm the source from which the initial attack originated. There is growing awareness of the possibility that cyber operations could lead to actual harm in the physical domain,[72] which would be equally true of cyber counter-attacks in self-defense.

Moreover, even if it is less likely, at least in the near term, that the use of autonomous cyber defenses will lead to human death, it is quite possible that a state's cyber defenses, pitted against another state's cyber algorithms, might ultimately place those two states in a situation of armed conflict. Given the potential that such a scenario could unfold as a result of

[71] Not all of these measures will necessarily amount to a use of force requiring a self-defense-based justification.

[72] *See* Nicole Perlroth & Clifford Krauss, *A Cyberattack in Saudi Arabia Had a Deadly Goal. Experts Fear Another Try.*, N.Y. TIMES, Mar. 15, 2018 (stating that goal of cyberattack on Saudi petrochemical plant was to trigger an explosion).

direct interaction between computer systems without an affirmative human decision, it may be wise to limit the capabilities of these systems. We recognise that a state's right of self-defense is applicable in the cyber sphere, and that in some circumstances this may necessitate the use of cyber counter-operations that respond faster than humans and with an ability to cause counter-damage. However, we suggest that any such operations be calibrated to have the lowest possible effect necessary in order to halt the attack against the system until a human can evaluate the situation, and that an alert system be put in place.[73] While self-defense may allow for operations that go beyond purely repelling the attack of the moment,[74] further counter-offensive operations in line with a wider understanding of the legal aims and limits of self-defense should only take place after consideration by the appropriate political and military leadership.

C. Transparency and Explainability

States generally are reluctant to disclose how they make decisions about resorting to force, and what information they have taken into account in making particular force-related decisions. One reason for this is that they wish to conceal their sources of information and the intelligence methods they employed to obtain that information. This lack of transparency may be amplified in contexts in which a state employs machine learning to assist it in making use of force decisions. For instance, the state may not be able to fully explain how or why the algorithm reached a particular conclusion, it may be loath to reveal the contents of the

[73] For instance, a state might create a trigger that alerts humans when one of its "automatic cyber defense" systems begins to undertake rapidly escalating exchanges with another system. A comparable context in which multiple algorithmic systems engaged in uncontrolled ways was the 2010 stock crash that was driven by the interaction of automatic trading systems. ECONOMIST, *One Big, Bad Trade*, Oct. 1, 2010.

[74] ILA Report, *supra* note 40.

algorithm, or it may even be reluctant to indicate that it deploys algorithms at all in the use of force context.

A significant challenge posed by the use of algorithms generally, and by machine learning in particular, is that individuals receiving the computer program's recommendations often will not know precisely how the program arrived at its conclusion. A common form of machine learning uses "neural nets," algorithms that recognize patterns by teaching themselves how to weight different components of the input data.[75] This weighting is often invisible. Thus, although the algorithm's creators will know the nature of the input data in general terms, they may not know the weight the algorithm gave to particular input points.[76] For instance, it may not be possible to determine the weight that an algorithm gave to troop movements, human intelligence, social media chatter, or indeed a previously unanticipated set of factors that the algorithm identified. Critics worry about reliance on algorithms that cannot explain to their users how they reached a particular outcome or recommendation;[77] these concerns may be compounded if states do not adequately address issues relating to algorithmic quality and reliability. A solution may take the form of "explainable algorithms," which are algorithms that can, to a greater or lesser extent, list the factors taken into account and explicate the weighting given to each factor.[78] Of course, even this task is not

[75] Robert Hof, *Deep Learning*, MIT TECH. REV. (2013).

[76] Danielle Kehl et al., *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, RESPONSIVE COMMUNITIES 28 (July 2017).

[77] Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV., Apr. 11, 2017.

[78] David Gunning, *Explainable Artificial Intelligence (XAI)*, Defense Advanced Research Projects Agency.

straightforward. The vast quantity of input data, and the complexity of the weighting and the relationships between data points, may pose challenges for human comprehension.[79]

Transparency and explainability are particularly important in a legal context, and various audiences often seek or demand greater transparency about state decision-making in the use of force context. A state may be pressed by the UN Security Council, its allies, other states, and its domestic and foreign publics to explain why it undertook a particular use of force. In limited cases, another state might even be able to pursue a case against it in the International Court of Justice, where it would have to account more clearly for its decision-making.[80] Further, being transparent about what information guided a decision to use force can help a state persuade other states that its force was legitimate (as happened with Israel's strike on the al Kibar facility in Syria). States therefore must be prepared to address questions about the use of algorithms in the *jus ad bellum* context, where other states and the public generally will be unsatisfied with a government explanation that asserts, "We went to war because the computer told us to."

In the area of threat assessments, in which states are more likely to use algorithms extensively, a greater form of transparency should, in principle, be achievable. A machine that simply provides a final conclusion that "an anticipated attack is estimated at 90%" is not transparent, and could lead to a high risk of automation bias in which the humans decide to act on the basis of percentages without understanding their origin. However, it should in theory be possible for the machine to reveal the information on which its estimation is based.[81] For example, it could note that its calculations are based on satellite imagery of troop

[79] Some have proposed that algorithms may provide a solution; algorithms may be designed to monitor the performance of other algorithms. Matt Burgess, *Holding AI to Account: Will Algorithms Ever Be Free From Bias If They're Created by Humans?*, WIRED, Jan. 11, 2016.

[80] Armed Activities on the Territory of the Congo (DRC v. Uganda), Merits [2005] ICJ Rep. 168.

[81] DARPA's pursuit of explainable AI is significant in this regard. *See* Gunning, *supra* note 78.

movements and the stock-piling of weapons by another state, combined with analysis of that state's past behaviour in similar circumstances. This would allow the executive and military leadership to make informed decisions, incorporate any additional contextual information of which they are aware, and identify what information might be missing from the machine's analysis.

Concerns over explainability and transparency are heightened in circumstances in which machines are designed to initiate action without human involvement (such as automated responses to cyber threats). If the deep learning process prevents the state from understanding how the machine's neural networks reach a conclusion, then in effect it becomes difficult—and perhaps impossible—to predict the actions a defensive system might take in response to an attack. Would it be lawful to deploy a system whose actions we cannot predict, and moreover to do so in the context of use of force? The answer may depend on how we define predictability and what it is that we need to predict. On the one hand, it might imply being able to predict each and every step the machine would take in any given context. This would allow for greater transparency and explainability, but would probably only be possible with systems that operate on the basis of pre-programmed rules of logic. On the other hand, if a state unleashed the full power of machine learning based on deep neural networks, it would likely mean that the state would be unable to predict each and every step the algorithm takes, because the algorithm would be able to adapt and develop new solutions to emerging and dynamic situations.

A different approach to predictability might be to accept the impossibility of knowing in advance the precise details of each step the machine might take, but require confidence that whatever actions it does initiate, the parameters of its programming ensure that these actions will always be within the limits of the applicable legal framework. In other words, the focus would be on reliability rather than predictability. This approach might be acceptable insofar as it does not allow for violations of the law on the resort to force, but it does create a serious challenge to the above-noted desire for transparency and explainability. If states choose to require the latter, it may prove impossible for now to use machine learning systems whose black box of reasoning we cannot crack. This concern is one more reason that, in the limited circumstances in which machines can initiate responses, they should be programmed to focus on the (more predictable) minimal defensive actions necessary, rather than allowing for counter-offensives.[82]

D. Attribution

It sometimes can be difficult for a victim state to correctly attribute the source of an armed attack. South Korea, for instance, took several weeks to attribute the sinking of one of its warships to North Korea.[83] Other examples include proxy wars during the Cold War era, where both the United States and the U.S.S.R. sought to conceal their participation in unconventional conflicts.[84] Nevertheless, it is imperative for a state to be able to make such an attribution to be able to know where, how, and against which actors to respond.

The process of attributing responsibility for a use of force to a specific state or entity is particularly difficult in a cyber context. This is due to both the nature of cyber operations and the architecture of the Internet. For example, cyber operations such as the Stuxnet worm may utilise malicious software, and it may be possible to identify that an attack is occurring,

[82] *See supra* Part III.B.

[83] Choe Sang-Hun, *North Korea Denies Sinking Navy Ship*, N.Y. TIMES, Apr. 17, 2010.

[84] Thomas Franck, *Who Killed Article 2(4)? Or: Changing Norms Governing the Use of Force by States*, 64 AM. J. INT'L L. 809, 820 (1970); Alberto Coll, *Unconventional Warfare, Liberal Democracies, and International Order*, 67 INT'L L. STUD. 3, 15 (1992) ("By its very nature, unconventional warfare leaves as few trails as possible. Conclusive, incontrovertible evidence of a party's guilt is hard to come by.").

and even that the attack is the result of particular software. Moving beyond this to attribute responsibility, however, gives rise to a number of complexities. For instance, determining the developer (or deployer) of malicious software is difficult. It may be possible to reverse engineer the software in order to extract the original source code, which can offer certain clues as to the author; it may be the case that the language used, or the repetition of previously identified code, provide concrete hints. It is equally possible, however, that the code's creators planted ostensible clues in the source code with the intent to mislead.

Moving beyond the strict confines of cyber, if an aggressor deploys a machine learning algorithm in the use of force context and the victim state is able to access that algorithm, the opacity inherent in the use of such algorithm may hinder a victim state's analysis of whether the attack was intentional or accidental. For instance, it may be difficult to understand the reasoning underpinning a machine learning recommendation or decision, and in any event, it may be straightforward for the aggressor to bury or disguise intent within the code. Of course, the "aggressor" state may still claim it was an accident, and the victim state may claim the right to respond regardless of intent,[85] but the use of machine learning presents increased opportunities for obfuscation.

Factors such as those described above render attribution difficult, and may limit the extent to which a victim state may exercise its right to self-defense. In particular, it may be the case that a victim state can take self-defense measures to halt an attack, but cannot take further, otherwise proportionate, measures because of its inability to directly attribute the attack. The complexity inherent in these issues may be illustrated by an example. Assume that a state suffers an armed attack that it tracks back to an algorithmic source but cannot attribute that algorithm to a recognized actor. Is it possible to speak of self-defense measures absent attribution of the armed attack? In past years there has been much discourse

[85] On the role of intent in the *jus ad bellum* see ILA Report, *supra* note 40.

surrounding self-defense against armed groups.[86] The United States, United Kingdom, and other states have argued that states may engage in self-defense against armed groups operating from the territory of other states but without attribution to the other state. That is, the right to self-defense is not predicated on the identity of the attacker but on the need of the victim state to use force to defend itself. However, in these scenarios, states and scholars have at least recognized the need to identify the entity behind the armed attack, even if it is an armed group and not a state, before considering a forcible response.

Taking the analogy from the realm of nonstate actors one step further, imagine that the source code appears to show that machines in State A have, through computer networks, caused significant material harm in State B, such as explosions in power stations and military installations that caused loss of life. State A denies responsibility and claims that it did not design or control the algorithm. State B has no evidence to disprove this claim, and it may well be that the algorithm is the work of a third state or a nonstate actor located in a third state. There is no love lost between States A and B, and despite the fact that this algorithm is operating from servers in the former, State A refuses to shut it down. If it cannot contain the damage through other means, State B might claim the right to defend itself by firing a missile at the server farm in the territory of State A, even without evidence of who is behind the attack. Essentially, the question raised by the potential use of untraceable armed attacks is whether the victim state that suffers an attack initiated by an algorithm outside its borders has a right to respond even if it cannot identify the entity—state or armed group—responsible for the algorithm. If so, states must consider whether additional limitations should regulate such a response in light of these special circumstances.

[86] *See, e.g.*, LUBELL, *supra* note 10, ch. 1-3; Michael Bothe, *Terrorism and the Legality of Pre-emptive Force*, 14 EUR. J. INT'L L. 227, 233 (2003); Sean Murphy, *Self-Defense and the Israeli Wall Advisory Opinion: An* Ipse Dixit *from the ICJ?*, 99 AM. J. INT'L L. 62, 64, 67-70 (2005).

E. Accountability for a Wrongful Use of Force

It is possible that the opacity associated with machine learning algorithms may frustrate accountability efforts. If a state makes a decision on the basis of machine learning, the state may not be able to identify or explain the reasoning underpinning that decision. As a result, that state may attempt to deflect scrutiny by pointing to the nature of algorithmic decision-making. A state in this position would effectively be arguing that, if it cannot foresee what an algorithm will do, it cannot be held responsible.

These suggestions fail to take into account existing obligations under international law. States deploy algorithms to perform particular tasks. As such, the state will be held responsible for the acts of the algorithm, in the same manner as it is held responsible for the acts of (human) state agents.[87] Responsibility cannot be negated simply by the decision to deploy an algorithm: this would undermine the entire framework of international law. Of course, states' continued responsibility poses particular challenges in the context of machine learning, but these are far from insurmountable, and a number of possibilities exist. States' due diligence obligations are particularly relevant. These require that states undertake a number of measures before deploying use of force algorithms, such as testing and impact assessments, and exercise continued oversight during their use. These measures would ensure that, although a state may not know the specific decision that an algorithm will reach in a particular situation, it must, at a minimum, assure itself as to the parameters within which the algorithm operates, the accuracy of the algorithmic model, its limitations, and whether it can make or simply inform a decision. While it is unreasonable to expect an error-free process, just as in any area of decision-making, it seems reasonable to demand a level of due

[87] This includes situations in which such agents act *ultra vires*.

diligence, according to which a state should be able to explain the algorithmic decision-making process and why it took a particular course of action.[88]

These requirements further indicate that the use of machine learning may actually facilitate transparency and accountability. Satisfying the above-mentioned international legal obligations before they deploy algorithms will force states to ensure that they have an appropriately structured decision-making process. This process, and (for instance) its emphasis on the need to understand how an algorithm weighted particular input factors, may enable a state to give an explanation of the reasoning underpinning a decision that is more accessible than human reasoning, and would be preferable to an explanation that simply invoked the official's "experience" or "intuition."

IV. Conclusion

The goal of this essay was to identify different decision-making points that arise in the context of states' use of force, to anticipate ways in which states may resort to machine learning algorithms to inform or make those decisions, and to highlight potential legal, policy, and ethical challenges that will arise as they do so. In general, we suggest that there may be situations in which the use of machine learning algorithms could actually improve the accuracy of states' use of force decision-making, but there are also situations in which states should be exceedingly cautious in resorting to machine learning tools.

There are some potentially broader implications for the use of force as machine learning tools proliferate, which we flag here for future consideration. One overarching

[88] As in non-algorithmic situations, the due diligence requirement focuses on the process rather than the outcome. As such, it is possible that an algorithm—just like a human—may "get it wrong" without giving rise to legal responsibility.

question is whether the gradual adoption of machine learning and, in the cyber context, the eventual automation of cyber responses will increase or decrease the likelihood that states will resort to force, offensively or defensively. Will machine learning tools further empower states that already have technologically sophisticated militaries and reduce certain existing barriers to resorting to force? Will these tools instead create a modest deterrent effect, at least as between state militaries that have each publicly deployed machine learning tools? Or will the ease of obscuring attribution using machine learning tools undermine the deterrent effect that Article 51 currently provides because victim states will be unable to assess where to direct their acts of self-defense?

There are other implications on the domestic plane, if we believe that an increasing use of algorithmic tools will widen the democracy deficit in national choices about resort to force (especially in the automatic cyber-defense context). Machine learning algorithms might disempower parliaments that have a constitutional role to play in decisions to go to war, and might, in view of their opacity, disempower interested publics who might oppose a given conflict. Now is the time, however, for these publics to make their views on machine learning known to the governments that will surely develop these tools.