View metadata, citation and similar papers at core.ac.uk brought to you by CORE
provided by Yale Law School Legal Scholarship Repository

Note

Autonomous Weapon Systems, International Crises, and Anticipatory Self-Defense

Nathan Leys[†]

I. INTRODUCTION

On the twelfth day of the Cuban Missile Crisis, the senior officers of the Soviet submarine B-59 faced a terrifying choice.[1] Nearby U.S. warships had begun dropping practice depth charges to force the submarine to surface, but the Russian sailors—who had been unable to contact Soviet command for two days—believed that the U.S. Navy was trying to kill them.[2] For their part, the Americans did not know that the B-59 was equipped with nuclear-tipped torpedoes, nor that the sub had prior authorization to use their nukes without direct permission from the Kremlin.[3] Making matters worse, the submarine's crew was exhausted and dangerously overheated—temperatures in the

[†] The author graduated from Yale Law School in May 2020. Profuse thanks are due to Rebecca Crootof, Paul Gewirtz, Rob Williams, and the talented editors of the Yale Journal of International Law.

[1] The details of the October 27, 1962 incident are drawn from Svetlana V. Savranskaya, *New Sources on the Role of Soviet Submarines in the Cuban Missile Crisis*, 28 J. STRAT. STUD. 233 (2005); Nicola Davis, *Soviet Submarine Officer Who Averted Nuclear War Honoured With Prize*, GUARDIAN (Oct. 27, 2017), https://www.theguardian.com/science/2017/oct/27/vasili-arkhipov-soviet-submarine-captain-who-averted-nuclear-war-awarded-future-of-life-prize.html.

[2] Savranskaya, *supra* note 1, at 245-49.

[3] Davis, *supra* note 1.

submarine had reached 45 degrees Celsius (113 degrees Fahrenheit).[4] Amid the stifling heat and the fear that World War III had already begun, the B-59's commanders had to decide whether to use their nuclear torpedoes in self-defense. The sub's second-in-command, Vasili Arkhipov, refused to give his consent because he believed that the U.S. was not attacking the ship, and therefore that a strike would not be justified as self-defense.[5] Under Soviet military protocols, his refusal meant the launch could not go forward.[6] The Cuban Missile Crisis ended the next day.

The events of October 27, 1962 highlight the importance of the legality of using force in self-defense in the context of an international crisis. But what if, instead of a handful of exhausted and terrified Soviet commanders, the decision to strike was made by a computer chip? In the next generation of crises, decisions such as these will be made by machines that think and act faster than Vasili Arkhipov could blink. Artificial intelligence is enabling a new generation of weapons that will dramatically change the nature of warfare.[7] The most capable—and notorious—of these are Autonomous Weapon Systems ("AWS"). Once activated, AWS can select and engage targets without specific human authorization.[8] This onboard selection, targeting, and firing capability is what distinguishes AWS from more rudimentary robotic weapons systems, such as the remotely piloted Predator drone; put differently, an autonomous weapon aims itself and pulls its own trigger.

[4] Savranskaya, *supra* note 1, at 246.

[5] *Id.* at 247.

[6] Davis, *supra* note 1.

[7] *See, e.g.*, PAUL SCHARRE, ARMY OF NONE: AUTONOMOUS WEAPONS AND THE FUTURE OF WAR 4 (2018) [hereinafter SCHARRE, ARMY OF NONE] ("Technology has brought us to a crucial threshold in humanity's relationship with war. In future wars, machines may make life-and-death engagement decisions all on their own. Militaries around the globe are racing to deploy robots at sea, on the ground, and in the air—more than ninety countries have drones patrolling the skies. These robots are increasingly autonomous and many are armed.").

[8] The question of how to define AWS is almost as hotly debated as their legality. *See, e.g.*, DEF. SCI. BOARD, SUMMER STUDY ON AUTONOMY 21 (June 2016) ("[AWS] upon activation can select and engage targets without human intervention."); Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837, 1854 (2015) (defining an AWS as "a weapon system that, based on conclusions derived from gathered information and preprogrammed constraints, is capable of independently selecting and engaging targets.") [hereinafter Crootof, *Killer Robots*]; Paul Scharre, *Autonomous Weapons and Operational Risk*, CTR. NEW AM. SEC. 3 (2016), https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515 [hereinafter Scharre, *Operational Risk*] (contrasting AWS as "weapons that would select and engage targets on their own" with "semi-autonomous weapons that would retain a human 'in the loop' for selecting and engaging specific targets."). *See generally* Michael Horowitz, *Why Words Matter: The Real World Consequences of Defining Autonomous Weapon Systems*, 30 TEMPLE INT'L & COMP. L.J. 85, 86 (2016) ("A challenge in the attempt to understand the ethical, legal, moral, strategic, and other issues associated with AWS is a basic lack of agreement over what an autonomous weapon actually is."). This overview is not intended to exhaust the definitional debate, but to emphasize the common theme of AWS' independence in selecting and engaging targets.

An example may clarify what we mean when we say a weapon is autonomous. Israel's military has commissioned the Harpy drone to search for and destroy enemy radars.[9] Once activated, this unmanned aerial vehicle (UAV) "can find targets autonomously based on radar or radio wave emissions."[10] Once the Harpy does so, it dive-bombs the radar installation, earning it the nickname the "kamikaze drone."[11] Physically, the Harpy is really not that different from a remote-controlled Predator drone. What makes the Harpies *autonomous* is their "ability to complete the engagement cycle—searching for, deciding to engage, and engaging targets—on their own."[12]

Autonomous systems are increasingly common: estimates vary, but one survey found over 280 weapons systems with some degree of autonomous capability, in all shapes and sizes.[13] Although the capabilities and deployment of these systems vary widely, the trend is clear: a substantial number of life-or-death decisions in the next generation of warfighting will be made by computers, not humans.[14] Nowhere are the stakes of these decisions higher than in an international crisis, in which the wrong move can spark a full-blown war. But although buckets of ink have been spilled over the implications of using AWS under *jus in bello* (the law governing conduct *during* a war),[15] remarkably little attention has been paid to the legal implications of AWS for *jus ad bellum* (the law of *starting* a war).[16] The question of whether AWS in the midst of a war will, for example, be able to adequately distinguish civilians from soldiers is incredibly important, and answering that question may help save innocent lives. But the best way to prevent collateral civilian casualties in a war is to prevent a war. AWS may make difficult issues of *jus ad bellum* harder; they may, in some cases, make them easier. But AWS *will* make *jus ad bellum* more complicated, and the lack of attention to this issue now risks the later unthinking application of pre-autonomy frameworks to a new reality when the stakes could not be higher, with all the potentially lethal frictions that entails.

[9] *See generally* SCHARRE, ARMY OF NONE, *supra* note 7, at 46-48; Tamir Eshel, *Drones Turned Into Weapons*, AVIATION WEEK (July 16, 2018), http://aviationweek.com/defense/drones-turned-weapons.html.

[10] Thomas Gibbons-Neff, *Israeli-Made Kamikaze Drone Spotted in Nagorno-Karabakh Conflict*, WASHINGTON POST (Apr. 5, 2016), https://www.washingtonpost.com/news/checkpoint/wp/2016/04/05/israeli-made-kamikaze-drone-spotted-in-nagorno-karabakh-conflict/?utm_term=.8727b5d61dc8.html.

[11] *Id.*

[12] SCHARRE, ARMY OF NONE, *supra* note 7, at 52.

[13] Heather Roff, *Weapons Autonomy is Rocketing*, FOREIGN POL'Y (Sept. 28, 2016), https://foreignpolicy.com/2016/09/28/weapons-autonomy-is-rocketing.html.

[14] SCHARRE, ARMY OF NONE, *supra* note 7, at 1-7.

[15] *See infra* notes 133, 137, 145 and accompanying text.

[16] I have found just two papers that meaningfully address how AWS may intersect with questions of *jus ad bellum* permissibility. *See* Ashley Deeks et al., *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 J. NAT'L SECURITY L. & POL'Y 1 (2019); Heather M. Roff, *Lethal Autonomous Weapons and Jus Ad Bellum Proportionality*, 47 CASE W. RES. J. INT'L L. 37 (2015) [hereinafter Roff, *Lethal Autonomous Weapons*].

The lack of attention to this issue does not only plague the commentariat. For example, the position papers submitted by the United States,[17] Russia,[18] and China[19] to the Convention on Certain Conventional Weapons' Group of Government Experts (CCW-GGE), the primary intergovernmental body addressing AWS, utterly fail to consider questions of *jus ad bellum*. So does the U.S. Department of Defense's seminal Directive 3000.09 on AWS.[20] This is an astonishing blind spot that risks dangerous consequences.

This Note is among the first to examine the legal questions raised by the use of force involving AWS during a crisis.[21] I take "crisis" here to mean a period of heightened tension between competing States, characterized by stepped-up military activities, but before either side has launched an armed attack against the other. Because the use of force outside the context of self-defense or UN Security Council authorization is always illegal under international law,[22] I ask instead under what conditions should an AWS be permitted to use force in anticipatory self-defense, and whether the actual or anticipated use of force by an AWS alters the legality of another party's self-defense.[23] Although the

[17] Submission from the United States to the Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons, Which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects, U.N. Doc. CCW/GGE.1/2017/WP.6 (Nov. 10, 2017) https://www.unog.ch/80256EDD006B8954/(httpAssets)/99487114803FA99EC12581D40065E90A/$file/2017_GGEonLAWS_WP6_USA.pdf [hereinafter Submission from the United States].

[18] Submission from the Russian Federation to the Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects, U.N. Doc. CCW/GGE.1/2018/WP.6 (Apr. 4, 2018), https://www.unog.ch/80256EDD006B8954/(httpAssets)/FC3CD73A32598111C1258266002F6172/$file/CCW_GGE.1_2018_WP.6_E.pdf.

[19] Submission from China to the Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which may be Deemed to be Excessively Injurious or to have Indiscriminate Effects, U.N. Doc. CCW/GGE.1/2018/WP.7 (Apr. 11, 2018), http://www.reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/GGE.1-WP7.pdf. The Chinese position paper does argue that "development and use of [AWS] would reduce the threshold of war, and the cost of warfare on the part of the user countries. This would make it easier and more frequent for wars to break out," but does not touch the question of whether and how AWS would affect the legality of the onset of such conflicts. *Id.* at 2.

[20] U.S. DEP'T OF DEF. DIR. 3000.09, AUTONOMY IN WEAPON SYSTEMS (Nov. 21, 2012) [hereinafter DoDD 3000.09]. The closest DoDD 3000.09 comes to addressing *jus ad bellum* is its generic statement that AWS must be employed "in accordance with the law of war," *id.* at 3, and its warning of "unintended engagements," *id.* at 1, 9-11.

[21] I use the formulation "involving AWS" because autonomous systems may be the actor, target, or some combination of the two.

[22] *See infra* notes 81-82 and accompanying text.

[23] There is, of course, substantial disagreement over whether the use of force in self-defense before suffering an armed attack is *ever* legal. *Compare* YORAM DINSTEIN, WAR, AGGRESSION AND SELF-DEFENCE 200 (5th ed. 2011) ("Regardless of the shortcomings of the [UN Charter] system, the option of a preventive use of force is excluded by Article 51.") *with* THOMAS M. FRANCK, RECOURSE TO FORCE: STATE ACTION AGAINST THREATS AND ARMED ATTACKS 98 (2002) ("Has recourse to such anticipatory

definitional boundaries of "anticipatory" self-defense are contested, I follow the commonly accepted *Caroline* standard: force may be used in self-defense "where the need to respond is 'instant, overwhelming, and leaving no choice of means, and no moment for deliberation.'"[24] I argue that with the right rules, the introduction of AWS into crises may improve compliance with *jus ad bellum* and prevent unintended escalation.

Such an inquiry is overdue for two reasons. First, there is a real prospect of an international crisis between the United States and a near-peer competitor in the foreseeable future.[25] AWS are likely to play an important role in such situations. Paul Scharre, one of the key authors of the U.S. Department of Defense's foundational policy on AWS,[26] argues that countries normally averse to deploying AWS are more likely to do so in crises as a way to mitigate the risk to their human soldiers in case of escalation.[27] However, "[t]he result . . . could be unintended escalation if the system engaged an otherwise legitimate enemy target but in a situation where the human operator did not intend an engagement.

self-defense in circumstances of extreme necessity been preserved, or repealed, by the Charter? Common sense, rather than textual literalism, is often the best guide to interpretation of international legal norms . . . . [N]o law—and certainly not Article 51—should be interpreted to compel the *reduction ad absurdum* that states invariably must await a first, perhaps decisive, military strike before using force to protect themselves."). As discussed below, this Note assumes the legality of anticipatory self-defense in at least some circumstances. *See infra* notes 81-92 and accompanying text.

[24] Ashley Deeks, *Taming the Doctrine of Pre-Emption*, *in* OXFORD HANDBOOK OF THE USE OF FORCE IN INTERNATIONAL LAW 662 n.5 (Marc Weller, ed., 2015) [hereinafter Deeks, *Taming the Doctrine of Pre-Emption*] (quoting Letter from Daniel Webster, U.S. Secretary of State, to Lord Ashburton, British Plenipotentiary (6 Aug. 1842), quoted in John Bassett Moore, *A Digest of International Law*, Vol. 2 (1906), § 217, at 412). This Note does not deal with cases in which the threat is less temporally immediate and the legality of using force is more open to question, which Professor Deeks places under the categories of "pre-emptive self-defense" and "preventive self-defense." *See also* TALLINN MANUAL 2.0 ON INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS 350 (Michael N. Schmidt ed., 2d ed. 2017) [hereinafter TALLINN MANUAL] (arguing for adopting the *Caroline* standard in the context of cyber armed attacks); *see infra* notes 84-87 and accompanying text (discussing the origins of the *Caroline* standard and competing definitions of pre-attack self-defense).

[25] *See, e.g.*, Peter Baker et al., *Seven Days in January: How Trump Pushed U.S. and Iran to the Brink of War*, N.Y. TIMES (Jan. 12, 2020), https://www.nytimes.com/2020/01/11/us/politics/iran-trump.html; Steven Lee Myers, *American and Chinese Warships Narrowly Avoid High-Seas Collision*, N.Y. TIMES (Oct. 2, 2018), https://www.nytimes.com/2018/10/02/world/asia/china-us-warships-south-china-sea.html (discussing tensions between the Chinese and U.S. navies in the South China Sea); Michael Birnbaum & Paul Sonne, *Trump's Ambassador to NATO Sets Off Diplomatic Incident With a Nuclear Edge*, WASH. POST (Oct. 2, 2018), https://www.washingtonpost.com/world/did-trumps-ambassador-to-nato-threaten-russia-with-preemptive-strikes/2018/10/02/243c5ba2-c664-11e8-9c0f-2ffaf6d422aa_story.html?utm_term=.1bfd23e38ef7 ("The U.S. ambassador to NATO set off alarm bells Tuesday when she suggested that the United States might 'take out' Russian missiles that U.S. officials say violate a landmark arms control treaty. . . . [A]rms control experts said [her comments] could be interpreted to mean a preemptive strike. Such a move could lead to nuclear war."). *See generally* Tim Swejis et al., *Back to the Brink: Escalation and Interstate Crisis*, HAGUE CTR. STRATEGIC STUD. 7 (2016), https://hcss.nl/sites/default/files/files/reports/HCSS_StratMon_Back_to_the_Brink.pdf ("For some time, many thought that interstate crisis was consigned to the dustbin of history. Even if it ever was, the lid has certainly come off in recent years . . . . Rather than isolated incidents, these events mark a larger trend: the comeback of interstate crisis.").

[26] DoDD 3000.09, *supra* note 20.

[27] Scharre, *Operational Risk*, *supra* note 8, at 22.

In a crisis, the consequences could be severe."[28] At the risk of oversimplification, it is impossible to know in advance how AWS will play a game of Chicken.[29] The legal issues of an AWS-involved first strike are an important part of a broader set of questions surrounding AWS and crisis decision-making.

Second, the legality of anticipatory strikes involving AWS will affect whether and how a crisis escalates.[30] There are three mechanisms by which this may occur. First, lawyers play an underappreciated role in operational decision-making.[31] If an anticipatory strike by an AWS would be prohibited or legally dubious, a professional military supplied with competent legal advice may choose not to deploy or activate such systems. Second, supporters of AWS contend that these systems may (eventually) be programmed to scrupulously follow all applicable laws and rules.[32] Assuming the possibility of such a system, an autonomous weapon's capacity to evaluate the legality of an anticipatory strike would be a crucial part of its programming. Third, whether an action is legal will affect how it is perceived by the policymakers and publics of each country involved in a crisis.[33] Public perception, in turn, plays a major role in

[28] *Id.*

[29] SCHARRE, ARMY OF NONE, *supra* note 7, at 209 ("Even a robot programmed to shoot only in self-defense could still end up firing in situations where humans wished it hadn't. If another nation's military personnel or civilians were killed, it might be difficult to de-escalate tensions.").

[30] This may be true whether or not the legal question is ever presented to an adjudicative body whose legitimacy and jurisdiction is generally accepted. *See*, *Legal Argumentation in International Crises: The Downing of Korean Air Lines Flight 007*, 97 HARV. L. REV. 1198, 1209-11 (1984) (arguing that international law shapes state behavior in crises even when there is no realistic possibility of sanctions for violations of international law).

[31] Rosa Brooks, a current Georgetown University Law Center professor and former Counselor to the Under Secretary of Defense for Policy, has discussed the role of lawyers in military decision-making. *See* ROSA BROOKS, HOW EVERYTHING BECAME WAR AND THE MILITARY BECAME EVERYTHING: TALES FROM THE PENTAGON 197-98 (2016) ("[The] U.S. military as a whole takes the laws of war very seriously. . . . Military lawyers undertake a wide variety of tasks . . . . [They] help commanders determine rules of engagement . . . and participate in real-time decision[-]making about targeting.").

[32] The primary advocate of this view is Ronald Arkin. *See, e.g.*, RONALD ARKIN, GOVERNING LETHAL BEHAVIOR IN AUTONOMOUS ROBOTS (2009); Ronald Arkin, *Lethal Autonomous Systems and the Plight of the Non-Combatant*, 137 AISB Q. 1 (2013); Ronald Arkin, *The Case for Ethical Autonomy in Unmanned Systems*, 9 J. MIL. ETHICS 332 (2010). The United States government adopted a similar view in its 2018 paper submitted to the Convention on Certain Conventional Weapons (CCW). *See* Submission from the United States, *supra* note 17.

[33] The January 2020 crisis in U.S.-Iranian relations is a case-in-point. Broadly speaking, arguments that the strike on Iranian Revolutionary Guard Corps Major General Qasem Soleimani was not justified by an imminent threat (or at least that the administration had failed to provide sufficient evidence of such a threat) formed the backbone of opposition to the strike and to further escalation. Julian E. Barnes et al., *Pressed for Details on Suleimani Strike, Trump Administration Gives Few*, N.Y. TIMES (Jan. 7, 2020), https://www.nytimes.com/2020/01/07/us/politics/trump-soleimani.html. At least in part as a result of the Trump administration's failure to gain widespread support for further escalation, *see* Nathaniel Rakich, *Americans Don't Know What to Think About Trump's Iran Strategy*, FIVETHIRTYEIGHT (Jan. 10, 2020), https://fivethirtyeight.com/features/americans-dont-know-what-to-think-about-trumps-iran-strategy, the United States did not respond in kind to a retaliatory missile strike on U.S. bases in Iraq that did not result in casualties. *See Iran Missile Attack: Did Tehran Intentionally Avoid U.S. Casualties?* BBC (Jan. 8, 2020), https://www.bbc.com/news/world-middle-east-51042156. Although it is impossible to know exactly what causal role legalistic arguments about the propriety of the Suleimani strike played in de-escalation, it seems clear that public perception of the legality of the strike was at least somewhat related to the U.S. decision not to escalate to a full-blown war.

whether a crisis escalates to war.[34] A public that sees itself as the victim of unlawful violence is more likely to be belligerent; a public that views its leaders' decision to take the country to war as illegal is more likely to be pacifistic.

Despite the importance of the AWS-*jus ad bellum* connection, the literature is worryingly silent on these issues. To be sure, the rise of AWS has sparked heated legal debates. Generally speaking, these discussions fall into three distinct but related categories. The first group of debates, following an influential 2012 report by Human Rights Watch and the Harvard Law School International Human Rights Clinic,[35] asks whether and when the development, deployment, and use of AWS could be permissible under the existing law of armed conflict (LOAC) or international humanitarian law (IHL).[36] The second set of debates addresses whether and how existing international law should be changed to address AWS, including the question of whether an international ban on such systems should be adopted.[37] The third category of legal debates regarding AWS examines how to apportion blame and liability should an autonomous weapon malfunction and/or harm noncombatants.[38]

[34] Perhaps the best-known authority for this proposition is ROBERT JERVIS, PERCEPTION AND MISPERCEPTION IN INTERNATIONAL POLITICS (1976). *See also* James D. Fearon, *Domestic Political Audiences and the Escalation of International Disputes*, 88 AM. POL. SCI. REV. 577 (1994) (discussing role of audience costs in crisis escalation) [hereinafter Fearon, *Domestic Political Audiences*]; James D. Fearon, *Rationalist Explanations for War*, 49 INT'L ORG. 379 (1995) (same) [hereinafter Fearon, *Rationalist Explanations*].

[35] *Losing Humanity: The Case Against Killer Robots*, HUM. RTS. WATCH (Nov. 19, 2012), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.html.

[36] *See, e.g.*, Robert Sparrow, *Twenty Seconds to Comply: Autonomous Weapon Systems and the Recognition of Surrender*, 91 INT'L L. STUD. 699 (2015); Gregory P. Noone & Diana C. Noone, *The Debate Over Autonomous Weapon Systems*, 47 CASE W. RES. J. INT'L L. 25 (2015); Kenneth Anderson & Matthew Waxman, *Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can*, HOOVER INST. (2013) https://media.hoover.org/sites/default/files/documents/Anderson-Waxman_LawAndEthics_r2_FINAL.pdf; Michael N. Schmitt & Jeffrey S. Thurnher, "*Out of the Loop:" Autonomous Weapon Systems and the Law of Armed Conflict*, 4 HARV. NAT'L SEC. J. 231 (2013); Tyler D. Evans, *At War with the Robots: Autonomous Weapon Systems and the Martens Clause*, 41 HOFSTRA L. REV. 697 (2013).

[37] *See, e.g.*, Mark Gubrud, *Why Should We Ban Autonomous Weapons? To Survive*, IEEE SPECTRUM (June 1, 2016), https://spectrum.ieee.org/automaton/robotics/military-robots/why-should-we-ban-autonomous-weapons-to-survive.html; Chris Jenks, *False Rubicons, Moral Panic, & Conceptual Cul-De-Sacs: Critiquing & Reframing the Call to Ban Lethal Autonomous Weapons*, 44 PEPP. L. REV. 1 (2016); John Lewis, *The Case for Regulating Fully Autonomous Weapons*, 124 YALE L.J. 1309 (2015); Anderson & Waxman, *supra* note 36.

[38] *See, e.g.*, Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 U. PA. L. REV. 1347 (2016) [hereinafter Crootof, *War Torts*]; Charles J. Dunlap, Jr., *Accountability and Autonomous Weapons: Much Ado About Nothing*, 30 TEMPLE INT'L & COMP. L.J. 63 (2016); Daniel N. Hammond, *Autonomous Weapons and the Problem of State Accountability*, 15 CHI. J. INT'L L. 652 (2015); Tim McFarland & Tim McCormack, *Mind the Gap: Can Developers of Autonomous Weapon Systems be Liable for War Crimes?*, 90 INT'L L. STUD. 361 (2014); *Mind the Gap: The Lack of Accountability for Killer Robots*, HUM. RTS. WATCH (Apr. 9, 2015), https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots.html.

A smaller but growing literature is beginning to address the potential strategic and political effects of AWS.[39] Some focus on the empirics and theory of domestic public opinion and AWS.[40] Others have reported on the competitive aspects of the race to develop AI-enabled weapons.[41] Still others—including the Author—have examined how the introduction of AWS might affect the escalation of international crises and strategic stability.[42] But amid this flurry of commentary, the dog that has not yet barked is the legality of first strikes when AWS are in the picture.

This Note proceeds in five parts. To highlight the stakes and illuminate the remainder of the analysis, Part II examines how AWS may affect how a crisis unfolds, with an emphasis on scenarios in which an AWS strikes first. Part III explores how AWS complicate the traditional *jus ad bellum* analysis. Part IV proposes safeguards to prevent AWS from sparking an illegal war, for example by requiring AWS, under certain circumstances, to absorb the first blow before shooting back. Part IV also briefly addresses enforcement. Part V concludes.

II. AUTONOMOUS WEAPONS AND INTERNATIONAL CRISES

Autonomous Weapon Systems will affect how and whether crises escalate. This is especially true for the most dangerous kinds of crises, involving two or more powerful and technologically advanced States. AWS capabilities are concentrated among a handful of countries with advanced militaries,

[39] *See, e.g.*, Andrew Massie, *Autonomy and the Future Force*, 10 STRAT. STUD. Q. 134 (2016).

[40] *See, e.g.*, Michael Horowitz, *Public Opinion and the Politics of the Killer Robots Debate*, 3 RES. & POL. 1 (2016); Frank Sauer & Niklas Schörnig, *Killer Drones: The 'Silver Bullet' of Democratic Warfare?*, 43 SEC. DIALOGUE 363 (2012); Charli Carpenter, *How Do Americans Feel About Fully Autonomous Weapons?*, DUCK OF MINERVA (June 19, 2013), http://duckofminerva.com/2013/06/how-do-americans-feel-about-fully-autonomous-weapons.html.

[41] The bulk of this literature focuses on the race between China and the United States to develop AWS. *See, e.g.*, Elsa B. Kania, *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*, CTR. NEW AM. SEC. (2017), https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/resources/docs/CNAS-Battlefield%20Singularity,%20Artifical%20intelligence,%20Military%20Revolution,%20and%20China's%20Future%20Military%20Power.pdf; Tom Upchurch, *How China Could Beat the West in the Deadly Race for AI Weapons*, WIRED (Aug. 8, 2018), https://www.wired.co.uk/article/artificial-intelligence-weapons-warfare-project-maven-google-china.html; John Markoff & Matthew Rosenberg, *China's Intelligent Weaponry Gets Smarter*, N.Y. TIMES (Feb. 3, 2017), https://www.nytimes.com/2017/02/03/technology/artificial-intelligence-china-united-states.html. I have written elsewhere that the United States desperately needs a sophisticated comparative analysis of how other countries view the role of AWS in their own and in the American militaries, similar to that of Dima Adamsky's famous comparative study of precision-guided munitions. Nathan Leys, *Autonomous Weapon Systems and International Crises*, 12 STRAT. STUD. Q. 48, 67-68 n.76 (2018); DIMA ADAMSKY, THE CULTURE OF MILITARY INNOVATION: THE IMPACT OF CULTURAL FACTORS ON THE REVOLUTION IN MILITARY AFFAIRS IN RUSSIA, THE UNITED STATES, AND ISRAEL (2010).

[42] *See, e.g.*, Leys, *supra* note 41; Jürgen Altmann & Frank Sauer, *Autonomous Weapon Systems and Strategic Stability*, 59 SURVIVAL 117 (2017); Heather Roff, *The Strategic Robot Problem: Lethal Autonomous Weapons in War*, 13 J. MIL. ETHICS 211 (2014).

including the United States, China, Russia, Israel, the U.K., and so on.[43] Although rudimentary autonomous capabilities may trickle down to less-advanced countries, the most capable—and therefore dangerous—AWS are likely to stay in the hands of those militaries between whom a war would be especially devastating.[44]

Legal debates around anticipatory self-defense and strategic debates over foreign policymaking in crises will draw on overlapping sets of facts. This is because the same facts that constitute a crisis will often provide the basis for a purportedly legal first strike. For example, troop movements near a disputed border are a common characteristic of international crises.[45] Those same troop movements may provide the factual predicate for a strike in anticipatory self-defense.[46] Although a comprehensive treatment of the mechanisms by which AWS might impact escalation between these countries is beyond the scope of this Note, this section briefly discusses four problems using facts abstracted from various real-life crises to illustrate some of the operational and legal consequences of introducing AWS into a crisis. These include the problem of false positives and trust in autonomous systems; spoofing and behavioral hacking; "flash crashes" resulting from unpredictable interactions between adversarial AWS that go south faster than humans can intervene; and AWS' ability to continue fighting when command-and-control networks have been damaged, but after hostilities have ceased.

[43] *See generally* Vincent Boulanin & Maaike Verbruggen, *Mapping the Development of Autonomy in Weapon Systems*, STOCKHOLM INT'L PEACE RES. INST. (2017), https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.

[44] *Cf.* Andrea Gilli & Mauro Gilli, *The Diffusion of Drone Warfare? Industrial, Organizational, and Infrastructural Constraints*," 25 SECURITY STUD. 67 (2016) (arguing that the proliferation of state-of-the-art unmanned vehicles has been less widespread and is more difficult than commonly assumed).

[45] In just the past few years, troop movements near disputed borders have characterized crises in Ukraine (*see* Michael Kofman et al., *Lessons from Russia's Operations in Crimea and Eastern Ukraine*, RAND CORP. (2017), https://www.rand.org/pubs/research_reports/RR1498.html), Gaza (*see* Anna Ahronheim, *IDF Beefs Up Troops in Gaza Border As Rioting Grows in the South*, JERUSALEM POST (Oct. 4, 2018), https://www.jpost.com/Arab-Israeli-Conflict/IDF-reinforces-southern-border-to-prevent-terror-from-Gaza-568645.html), India-Pakistan (*see* Jeffrey Gettleman, *Troops on the Move as India And Pakistan Face Off Over Kashmir*, N.Y. TIMES (Feb. 25, 2019), https://www.nytimes.com/2019/02/25/world/asia/kashmir-india-pakistan-pulwama.html), and the Korean Peninsula (*see* Choe Sang-Hun, *North Korea Renews Guam Threat Ahead of Joint Naval Exercise*, N.Y. TIMES (Oct. 13, 2017), https://www.nytimes.com/2017/10/13/world/asia/north-korea-renews-guam-threat-ahead-of-joint-naval-exercise.html).

[46] *See, e.g.*, Miriam Sapiro, *Iraq: The Shifting Sands of Preemptive Self-Defense*, 97 AM. J. INT'L L. 599, 601 (2003) ("In 1967, Israel launched an attack on the Egyptian army massing on its borders . . . . [T]his action is frequently cited as the classic modern case of legitimate anticipatory self-defense."). One scholar has noted that Israel initially attempted to characterize its strike as "an act of self-defense taken in response to an Egyptian attack" that had already begun, rather than as an act of anticipatory self-defense. David A. Sadoff, *A Question of Determinacy: The Legal Status of Anticipatory Self-Defense*, 40 GEO. J. INT'L L. 523, 567 (2009).

A. *The Canadian Geese Problem: False Positives and Trust*

The AI-driven increase in the tempo of battle means that soldiers will be under intense time pressure to either authorize action or stand down when faced with an autonomous system's (possibly incorrect) warning of an incoming attack. In the mid-1950s, a U.S. early warning system picked up what it believed to be a Soviet missile launch. Fortunately for the United States, the "projectile" turned out to be a flock of geese. Fortunately for the U.S.S.R., someone in the U.S. chain-of-command apparently had the presence of mind not to press the proverbial red button. In the waning days of the Cold War, Mikhael "Gorbachev kept a sculpture of a goose in his Moscow office," a reminder of the danger of placing too much trust in a computer's interpretation of a crisis.[47]

But a soldier presented with a warning of an inbound attack today may have no choice but to trust her computer. "Just as autonomy is needed to successfully defend against saturation attacks from missiles and rockets today, it similarly might be needed in other future [combat] situations where human reaction times are too slow to be successful."[48] For example, putting a sailor in charge of manning a warship's defenses may doom the whole crew if there are too many missiles moving too quickly for a human to shoot down—but putting a computer in charge of the anti-missile system may save the ship. And as AWS (and AI generally) advance, that soldier may become predisposed to "automation bias," in which "people presume that algorithmic decisions are made on the basis of indisputable hard science, or operate at a level beyond human capacity, or . . . [they] fear overruling the computer and 'getting it wrong.'"[49] Compounding this phenomenon, militaries may endeavor to increase their soldiers' trust in AWS in an effort to improve the effectiveness of human-robot teams.[50]

A soldier's trust in an AWS, born of necessity and training, may be desirable in many—even most—cases. But when an AWS does not require human authorization to engage, when events move so quickly that a soldier does not have the chance to perform a threat analysis separate from the machine's and is forced to defer to the computer, or when a soldier trusts a machine more than she trusts her own judgment, false alarms can have disastrous consequences. The

[47] DAVID E. HOFFMAN, THE DEAD HAND: THE UNTOLD STORY OF THE COLD WAR ARMS RACE AND ITS DANGEROUS LEGACY 475 (2009).

[48] Scharre, *Operational Risk*, *supra* note 8, at 46.

[49] Deeks et al., *supra* note 16, at 18.

[50] Heather M. Roff & David Danks, *"Trust but Verify": The Difficulty of Trusting Autonomous Weapon Systems*, 17 J. MIL. ETHICS 2 (2018).

Patriot fratricide debacle during the Iraq War is illustrative.[51] During the 2003 invasion of Iraq, the United States employed the Patriot missile-defense system to shoot down Iraqi Scud missiles. But on three different occasions, Patriot batteries mistook U.S. or British jets for hostile targets. In its autopsy of the friendly fire incidents, the DOD's Defense Science Board concluded that "the Patriot system . . . [was] a poor match to the conditions of [Operation Iraqi Freedom]. The operating protocol was largely automatic, and the operators were trained to trust the system's software."[52] Imagine if, during a crisis but before the outbreak of violence, a missile-defense battery operating near a disputed border mistook a rival's routine air patrol (or a flock of Canadian geese) for an incoming missile. Systems like the Patriot do not only risk friendly-fire in a war; they also risk starting one.

B. *Hell Hath No Fury Like a Tricked Robot: Spoofing and Behavioral Hacking*

A second problem created or exacerbated by introducing AWS into crises is that of spoofing or behavioral hacking. These terms describe an adversary's efforts to trigger an undesirable reaction from an artificially intelligent system.[53] This is similar to the Canadian Geese problem, except the false positive is fed to the AWS by an adversary or third party. In the context of anticipatory strikes by AWS, this could involve an adversary or third party feeding an autonomous weapon information designed to cause the AWS to fire first.

Recall the Harpy, Israel's anti-radar "kamikaze drone." The IDF's perennial adversaries, Palestinian militant groups like Hamas, "have been using human shields, hospitals, schools, UN facilities, mosques, hotels and private homes to hide and protect personnel and equipment since the late 1960s."[54] In the 2014 Gaza conflict, Hamas fired rockets from these locations "to provoke

[51] A *60 Minutes* exposé on the fratricides explained, "if the Patriot displays the symbol for an incoming ballistic missile, its operator has just seconds to decide whether to override the machine, or let it fire." Rebecca Leung, *The Patriot Flawed? Failure to Correct Problems Led to Friendly Fire Deaths*, 60 MINUTES (Feb. 19, 2004), https://www.cbsnews.com/news/the-patriot-flawed-19-02-2004.

[52] *Report of the Defense Science Board Task Force on Patriot System Performance: Report Summary*, OFF. UNDER SEC'Y DEF. ACQUISITION, TECH. & LOGISTICS 2 (2005), https://dsb.cto.mil/reports/2000s/ADA435837.pdf.

[53] Paul Scharre, *The Lethal Autonomous Weapons Governmental Meeting (Part I: Coping with Rapid Technological Change)*, JUST SECURITY (Nov. 9, 2017), https://www.justsecurity.org/46889/lethal-autonomous-weapons-governmental-meeting-part-i-coping-rapid-technological-change.html ("Deep neural networks . . . have shown a particular vulnerability to a form of spoofing attack where the machine is fed false data to manipulate its decision-making. . . . [T]hese spoofing attacks can be hidden so that they are invisible to humans, and there is currently no known effective defense against this form of attack.").

[54] Eitan Shamir & Eado Hecht, *Gaza 2014: Israel's Attrition vs Hamas' Exhaustion*, 44 PARAMETERS 81, 85 (2014).

retaliatory fire."[55] It is entirely conceivable that rather than firing rockets from a school, which may strengthen the legitimacy of a self-defense claim by Israel, Hamas could instead place a targeting radar on top of the same school and "spoof" a Harpy into firing the first shot.

C. *The "Flash Crash" Fear: Competing Algorithms and Catastrophic Interaction*

The interaction of two adversarial Autonomous Weapon Systems in a tense but non-violent situation poses the risk of unintentional escalation at dizzying speeds, otherwise known as a "flash crash". Although such a scenario is not known to have occurred at the time of writing, other examples of negative interactions between competitive, algorithm-based systems do not provide much reason for comfort. For instance, "automated stock trading algorithms offer an example of the risks of autonomous systems interacting in complex, competitive environments and at speeds exceeding human reaction times."[56] In 2010, competing algorithms set off a "flash crash" that wiped out around 10 percent of the Dow Jones Industrial Average in a matter of minutes.[57] Outside the stock market context, two competing buy-sell algorithms on Amazon bid the price of an obscure textbook on flies up to $23.7 million dollars (plus $3.99 shipping).[58] The fear is that two militaries' AWS will interact in a way that creates a rapid feedback loop with destructive consequence.[59]

The risk of unforeseeable adversarial interaction is exacerbated by the possibility that an AWS may act to achieve its goals in a way that humans cannot predict or understand. The experience of AlphaGo is instructive.[60] Go is an ancient and incredibly complex game of strategy. In 2017, an AI called AlphaGo, which had been trained on sets of expert human games, defeated the world's top human Go player. Later, DeepMind—a research offshoot of Google and the creator of AlphaGo—released recordings of a new AI, AlphaGo Zero, which had been trained only by playing itself.[61] The self-trained program defeated the human-trained AI, 100-0.[62] Go experts described AlphaGo Zero's strategies as

[55] *Id.*
[56] Scharre, *Operational Risk*, *supra* note 8, at 35.
[57] *Id.*
[58] Olivia Solon, *How A Book About Flies Came to Be Priced $24 Million On Amazon*, WIRED (Apr. 27, 2011), https://www.wired.com/2011/04/amazon-flies-24-million/.
[59] Ulrike Esther Franke, *Flash Wars: Where Could An Autonomous Weapons Revolution Lead Us?*, EUR. COUNCIL FOREIGN REL. (Nov. 22, 2018), https://www.ecfr.eu/article/Flash_Wars_Where_could_an_autonomous_weapons_revolution_lead_us ("With [AWS], 'flash crashes' could turn into 'flash wars'.").
[60] *See* Leys, *supra* note 41, at 53.
[61] David Silver et al., *Mastering the Game of Go Without Human Knowledge*, 550 NATURE 354, 354 (2017).
[62] *Id.*

"moments of algorithmic inspiration," "previously unknown," and "creativ[e]."[63] As the DeepMind researchers write, training AIs on *human* strategies "may impose a ceiling on … performance."[64] But training AIs to break that ceiling may make the AI's actions incomprehensible to humans. Assuming that comprehensibility and performance are at least somewhat inversely related and that no two countries' AWS will be trained in exactly the same way,[65] it is hard to know how human soldiers will interpret an AWS's actions. And it is harder still to predict how those actions will be interpreted by an adversarial AWS.

D. *The "Battle of New Orleans" Problem: AWS and Disaggregated Command-and-Control* [66]

AWS' ability to fight when disconnected from their handlers is both a feature and a bug, at least when hostilities were once ongoing but have since ceased. Conceptually, this problem is not new. On January 8, 1815, British forces attacked American troops under the command of Andrew Jackson during the Battle of New Orleans.[67] The clash occurred during peacetime; unbeknownst to the combatants, the United States and the United Kingdom had already signed the Treaty of Ghent, ending the War of 1812.[68] Of course, Andrew Jackson did not have access to a satellite phone.[69] The last two centuries have seen dramatic

[63] Larry Greenemeir, *AI versus AI: Self-Taught AlphaGo Zero Vanquishes Its Predecessor*, SCI. AM. (Oct. 18, 2017), https://www.scientificamerican.com/article/ai-versus-ai-self-taught-alphago-zero-vanquishes-its-predecessor.

[64] Silver et al., *supra* note 61, at 354.

[65] Scharre, *Operational Risk*, *supra* note 8, at 36 ("[C]ompetitors are not likely to share their algorithms with one another.").

[66] Professor Rebecca Crootof has helpfully pointed out to me that this phenomenon might be more accurately described as a problem for *jus post bellum* (which roughly translates as "the law of ending war") rather than *jus ad bellum*. *See generally* Carsten Stahn, Jus Post Bellum*: Mapping the Discipline(s)*, 23 AM. U. INT'L L. REV.311 (2008); Carsten Stahn, 'Jus Ad Bellum', 'Jus In Bello' . . . 'Jus Post Bellum'? – *Rethinking the Conception of the Law of Armed Force*, 17 EUR. J. INT'L L. 921 (2006). To the extent that *jus post bellum* is a legal concept with independent force (rather than a philosophical outgrowth of just war theory), it has largely been developed in the context of a major power's and the international community's responsibilities to the people of an occupied state after the defenestration of the latter's government. *See, e.g.*, Dana Wolf, *Transitional Post-Occupation Obligations Under the Law of Belligerent Occupation*, 27 MINN. J. INT'L L. 5 (2018). Where AWS have a post-war policing role, I agree that a discussion of *jus post bellum* and autonomy is important. *Cf.* Christof Heyns, *Human Rights and the Use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement*, 38 HUM. RTS. Q. 350 (2016). Such a discussion is, of course, beyond the scope of this Note. In my view, *jus post bellum* as it has been developed in the last two decades has little applicability to a situation in which a military-to-military clash breaks a ceasefire between two states of comparable power and ignites a new conflict; *jus ad bellum* governs the recourse to force when hostilities are not occurring, regardless of whether hostilities have occurred in the recent past, and therefore governs here.

[67] Abraham D. Sofaer, *Emergency Power and the Hero of New Orleans*, 2 CARDOZO L. REV. 233, 240 (1980).

[68] *Id.* at 241.

[69] A similar situation (strategically speaking) confronted the Japanese holdouts, or *Zanryū Nipponhei*, Japanese Imperial soldiers stationed on remote Pacific islands who were unaware (or did not believe)

improvements in the "command-and-control" (C2) structures on which modern military commanders rely to collect information from, and relay orders to, troops in the field. But the modern communications networks on which the United States, its allies, and its peer/near-peer competitors rely may well be targeted early on in a conflict.[70] AWS will be especially valuable in degraded C2 environments; unlike Predator drones, for example, their ability to fight does not depend on the quality or even existence of a satellite uplink to an Air Force base in Nevada.[71] In addition, autonomy may reduce the risk of hacking[72] and the strain on C2 networks even in the best of times.[73]

DARPA's Collaborative Operations in Denied Environments (CODE) program illustrates concretely how AWS might function when they cannot contact their human operators.[74] CODE's purpose is "to design sophisticated software that will allow groups of drones to work in closely coordinated teams, even in places where the enemy has been able to deny American forces access to GPS and other satellite-based communications."[75] Although CODE's purpose is not explicitly to develop fully autonomous weapons, it does seek to leverage autonomy to reduce the need for direct human control of unmanned systems.[76] The strategic and technical goals of CODE are a short step from realizing AWS that can function in denied environments.[77]

that WWII had ended. *See, e.g.*, Justin McCurry, *Hiroo Onoda: Japanese Soldier Who Took Three Decades to Surrender, Dies*, GUARDIAN (Jan. 17, 2014), https://www.theguardian.com/world/2014/jan/17/hiroo-onoda-japanese-soldier-dies. *See also* Leys, *supra* note 41, at 58, 66 (discussing the example of the *Zanryū Nipponhei*).

[70] *See, e.g.*, Andrew Massie, *Autonomy and the Future Force*, 10 STRAT. STUD. Q. 134, 146 (2016) ("If the adversaries we expect to face take the battlefield, the long screw driver will be consigned to history . . . ."); James Dobbins, *War With China*, 54 SURVIVAL 7, 15 (2012) ("Chinese cyber and anti-satellite capabilities may in time be able to disrupt US C4ISR [command, control, communications, computers, intelligence, surveillance, and reconnaissance] capabilities . . . .").

[71] *See* W.J. Hennigan, *Drone Pilots Go to War in the Nevada Desert, Staring at Video Screens*, L.A. TIMES (June 17, 2015), https://www.latimes.com/nation/la-na-drone-pilots-20150617-story.html.

[72] On the risk of an AWS being hacked, *see* Michal Klincewicz, *Autonomous Weapon Systems, the Frame Problem and Computer Security*, 14 J. MIL. ETHICS 162 (2015).

[73] Daniel Gonzales & Sarah Harting, *Designing Unmanned Systems with Greater Autonomy: Using a Federated, Partially Open Systems Architecture Approach*, RAND CORP. (2014) at xii, http://www.rand.org/content/dam/rand/pubs/research_reports/RR600/RR626/RAND_RR626.pdf ("[A]utonomous functions [may] reduce messaging loads on communications links to C2 and information analysis centers. For example, autonomous onboard planning algorithms can help reduce communications loads and lessen the need for frequent maneuver, heading, or flight commands.").

[74] *See generally* Scott Wierzbanowski, *Collaborative Operations in Denied Environment (CODE)*, DEF. ADVANCED RES. PROJECTS AGENCY, https://www.darpa.mil/program/collaborative-operations-in-denied-environment (last visited May 21, 2020). CODE is not the only U.S. program seeking to realize the military benefits of AI-enabled systems. *See* Boulanin & Verbruggen, *supra* note 43, at 94-97.

[75] Kelsey D. Atherton, *Are Killer Robots the Future of War? Parsing the Facts on Autonomous Weapons*, N.Y. TIMES (Nov. 15, 2018), https://www.nytimes.com/2018/11/15/magazine/autonomous-robots-weapons.html.

[76] *Id.*

[77] *See, e.g.*, Jamie Condliffe, *A 100-Drone Swarm, Dropped from Jets, Plans Its Own Moves*, MIT TECH. REV. (Jan. 10, 2017), https://www.technologyreview.com/2017/01/10/154651/a-100-drone-swarm-dropped-from-jets-plans-its-own-moves/ (discussing successful DOD testing of a supervised autonomous UAV swarming concept).

Now imagine if an AWS, severed from its C2 network by accident, attack, or design,[78] were forced to decide whether to engage a nearby target.[79] For example, MK 60 CAPTOR (encapsulated torpedo) mines "detect and classify submarines and release a modified torpedo" to attack enemy targets.[80] If such an autonomous torpedo launcher, stationed in a crucial shipping lane during a conflict and cut off from C2 before the declaration of a ceasefire, picked up an adversary's warship bearing down on it, such a weapon might—like Andrew Jackson's forces at New Orleans—decide to attack under the mistaken assumption that hostilities were ongoing. Such an attack might well scuttle peace talks and erase the credibility of one party's promise to hold its fire.

III. ANTICIPATORY SELF-DEFENSE AND AUTONOMOUS WEAPON SYSTEMS

Under the U.N. Charter, a State may legally use force in only two scenarios: self-defense and pursuant to authorization by the U.N. Security Council.[81] This Note deals with the former, as the collective security provisions of the U.N. Charter are effectively defunct.[82] Furthermore, it is difficult to imagine how the use of AWS in a Security Council-authorized action would

[78] *See, e.g.*, Jeremy Straub, *Consideration of the Use of Autonomous, Non-Recallable Unmanned Vehicles and Programs as a Deterrent or Threat by State Actors and Others*, 44 TECH. SOC. 39 (2016).

[79] One might also draw a comparison to the crew of the nuclear submarine USS *Alabama* in the 1995 film *Crimson Tide*. The *Alabama* receives an order to launch its nuclear missiles against a Russian military base fueling its ICBMs, but before launching begins to receive another message from U.S. Strategic Command. The second message is interrupted in dramatic fashion by the attack of a Russian submarine, and after defeating the Russian sub the *Alabama*'s captain and second-in-command must decide whether to continue with the launch or hold off until contact with Strategic Command can be reestablished. Needless to say, the film concludes with Denzel Washington saving humanity. CRIMSON TIDE (Hollywood Pictures 1995).

[80] *MK 60 Encapsulated Torpedo (CAPTOR)*, GLOBAL SEC., https://www.globalsecurity.org/military/systems/munitions/mk60.htm. *See also* Scott C. Truver, *Naval Mines and Mining: Innovating in the Face of Benign Neglect*, CTR. INT'L MARITIME SEC. (Dec. 20, 2016), http://cimsec.org/naval-mines-mining-innovating-face-benign-neglect/30165 (discussing the potential role of next-generation naval mines in the Third Offset).

[81] U.N. Charter art. 2, ¶ 4 ("All Members shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state, or in any other manner inconsistent with the Purposes of the United Nations."); U.N. Charter art. 42 (discussing use of force pursuant to Security Council authorization); U.N. Charter art. 51 ("Nothing in the present Charter shall impair the inherent right of individual or collective self-defence if an armed attack occurs against a Member of the United Nations . . . .").

[82] As Professor Franck pointed out nearly half a century ago, "[a]lmost from the moment the San Francisco Charter was signed, [the] essential prerequisite for U.N. collective enforcement action—the unanimity of the great Powers [on the Security Council]—was seen to be an illusion. . . . As Chapter VII [collective security] was seen to rust, increasing use began to be made of Articles 51, 52, and 53, which set out the rights of states" to self-defense. Thomas M. Franck, *Who Killed Article 2(4)? or: Changing Norms Governing the Use of Force by States*, 64 AM. J. INT'L L. 809, 810-11 (1970); *see also* David Kaye, *Adjudicating Self-Defense*, 44 COLUM. J. TRANSNAT'L L. 134, 148 (2005) ("The UN Charter, with the Security Council as its enforcer, was drafted on the premise that the UN's collective security machinery would *work*. But since that has often not been the case, reliance by states on self-defense has been the norm. . . . [S]elf-defense is not the exceptional right imagined by some; it is a typical way in which states enforce security.").

render an otherwise permissible use of force unlawful under the *jus ad bellum* inquiry. This Note does not address whether an autonomous system could legally use force in self-defense once a conflict has begun, an inquiry that belongs under *jus in bello*.

This Note instead asks how the introduction of AWS changes the inquiry into when States may lawfully make the transition between peace and war, or, put differently, between crisis and violence. Specifically, under what conditions may AWS use force in anticipatory self-defense?

Much ink has been spilled over whether anticipatory strikes can ever be legal under international law, a question which generally turns on how a given author defines "anticipatory" self-defense.[83] As noted above, I follow the modern consensus in defining anticipatory self-defense in line with the *Caroline* Doctrine, which permits the use of force when the party claiming anticipatory self-defense can "show a necessity of self-defense, instant, overwhelming, leaving no choice of means, and no moment for deliberation."[84] Professor Deeks distinguishes this from "preemptive self-defense," meaning "the use of force in self-defence to halt a particular tangible course of action that the potential victim State perceives will shortly evolve into an armed attack against it," and "preventive self-defense," meaning "the use of force in self-defence to halt a serious threat of an armed attack, without clarity about when or where that attack may emerge."[85] For these latter two categories, in which the threat of armed attack is less temporally immediate, the validity of claims of self-defense is generally acknowledged to be weaker, which is why I do not address them here.[86]

The applicability of the *Caroline* standard to the brave new world of AWS finds support in the law of nuclear weapons. Speed, rather than destructive potential, distinguishes both nuclear weapons and AWS from other conventional weapons. In its 1996 Advisory Opinion on the Legality of the Threat or Use of

[83] *See generally* Sean D. Murphy, *The Doctrine of Preemptive Self-Defense*, 50 VILL. L. REV. 699 (2005). *See also* DINSTEIN, *supra* note 23, at 204-05 (arguing that the U.N. Charter forecloses claims of self-defense that are "conjectural" in any measure, but allows for "interceptive" self-defense when an enemy has "committed itself to an armed attack in an ostensibly irrevocable way."); FRANCK*, supra* note 23, at 107 (asserting that "States seem willing to accept strong evidence of the imminence of an overpowering attack as tantamount to the attack itself, allowing a demonstrably threatened state to respond under Article 51 as if the attack had already occurred, or at least to treat such circumstances, when demonstrated, as mitigating the system's judgment of the threatened state's pre-emptive response.").

[84] *See supra* note 2 and accompanying text; Deeks, *Taming the Doctrine of Pre-Emption*, *supra* note 24. *See also* Michael Wood, *The* Caroline *Incident*, *in* THE USE OF FORCE IN INTERNATIONAL LAW: A CASE-BASED APPROACH 5-14 (Tom Ruys et al. eds., 2018); James A. Green, *Docking the* Caroline*: Understanding the Relevance of the Formula in Contemporary Customary International Law Concerning Self-Defense*, 14 CARDOZO J. INT'L & COMP. L. 429 (2006).

[85] Deeks, *Taming the Doctrine of Pre-Emption*, *supra* note 24, at 662-63.

[86] *Id.*

Nuclear Weapons, the ICJ suggested that the use of nuclear weapons in anticipatory self-defense may be permissible "in an extreme circumstance of self-defence, in which [a State's] very survival would be at stake."[87] But as Thomas Schelling famously noted, even with the advent of the nuclear weapons, "[i]t is not true that for the first time in history man has the capability to destroy a large fraction, even the major part, of the human race. . . . Against defenseless people there is not much that nuclear weapons can do that cannot be done with an ice pick."[88] The fundamental contribution of nuclear weapons "is not in the number of people they can eventually kill but in the speed with which it can be done. . . . [N]uclear weapons make it *possible* to compress the fury of global war into a few hours."[89] AWS provide a similar contribution, greatly accelerating the tempo of battle.[90] Of course, the prototypical AWS—*e.g.*, a close-in ship defense system like the U.S. Navy's Phalanx CIWS or a swarm of small, expendable drones—is far less destructive than a nuclear weapon. But because *speed* is what distinguishes nuclear weapons, the logic of the 1996 Advisory Opinion translates rather well to AWS.

Resolving the question of whether anticipatory self-defense can ever be legal lies well beyond the scope of this Note. This Note, like most experts and the U.S. foreign policy establishment,[91] assumes that anticipatory strikes that move a situation from crisis to violence are legal when they meet the *Caroline* standard (or something close to it).[92] But the mechanical application of static *jus ad bellum* principles to AWS produces certain tensions and may even lead to unwanted escalation. In this Part, I present five legal friction points of AWS-involved anticipatory strikes: the evidentiary burden a State must meet to make out a claim of self-defense under international law; whether the standard for evaluating a State's self-defense claim is objective, subjective, or some combination of the two; the question of mistaken self-defense; the requirement of *jus ad bellum* proportionality (as distinct from its *jus in bello* counterpart); and the issue of whether and when an AWS's use of force qualifies as an "armed attack" giving rise to a right of self-defense.

[87] Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, ¶ 97 (July 8).

[88] THOMAS SCHELLING, ARMS AND INFLUENCE 19 (1966).

[89] *Id.* at 20-21 (emphasis in original).

[90] Paul Scharre, *A Million Mistakes a Second*, FOREIGN POL'Y (Sept. 12, 2018), https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/.

[91] *See, e.g.*, AIDEN WARREN & INGVILD BODE, PREVENTIVE AND PREEMPTIVE SELF-DEFENSE IN US NATIONAL SECURITY POLICY: A BRIEF HISTORY 46-59 (2014) (arguing that U.S. foreign policy has long contemplated anticipatory self-defense); Anthony Clark Arend, *International Law and the Preemptive Use of Military Force*, 26 WASH. Q. 89 (2003) (arguing that anticipatory self-defense "does not violate international law because the [U.N.] charter framework is no longer reflected in state practice" and therefore is no longer good law.).

[92] *See also* Deeks, *Taming the Doctrine of Pre-Emption*, *supra* note 24, at 662 (describing anticipatory self-defense as the "least-controversial form of pre-attack self-defense . . . The view that a state must wait to suffer an armed attack before being able to respond forcibly now appears to be a minority view.").

A. *Evidentiary Burdens*

A self-defense claim requires factual support. The introduction of AWS will make the collection and presentation of the evidence of an armed attack giving rise to a right to self-defense easier in some cases, but harder in others. As the ICJ reiterated in the *Oil Platforms* case, "the burden of proof of the facts showing the existence of such an attack [justifying the use of force in self-defense] rests on the" party claiming self-defense.[93] In many—perhaps most—cases, AWS will not make this task more difficult than it already was. Indeed, if AWS are equipped with sensor capabilities beyond that of a human, their recordings of events may even make it easier for a State to prove its self-defense claim after the fact.[94]

In the case of expendable AWS that are not in constant contact with a command system, however, claims of self-defense may be all-but-impossible to prove. Take so-called "fire and forget" missiles, like the UK's Brimstone missile or the aforementioned Israeli Harpy drone. These weapons need not communicate with a centralized command structure and are valuable because they are "relatively inexpensive . . . and expendable."[95] Consider again the spoofing example of a Hamas radar installation placed on top of a school. Normally, if a pilot detects a target-lock by an enemy missile, the pilot may pull the trigger first in a justifiable use of self-defense because an "an armed attack [against the pilot] may be deemed to be in progress."[96] But if an Israeli Harpy detects a Hamas radar lock and decides to strike, and the data recorded by the drone have not also been recorded by some other Israeli unit, all the data that the drone used to make that decision will be destroyed in the subsequent dive-bombing. Israel would find it impossible to meet its burden of proof, because the act of striking would destroy the evidence necessary to justify the strike.

B. *An Objective or Subjective Standard?*

The *Caroline* standard is straightforward enough, but on its own does not answer the question: instant and overwhelming need according to whom? To justify its use of force against State Z, State X has the burden of proof to show

[93] Case Concerning Oil Platforms (Iran v. U.S.), 2003 I.C.J. 161, ¶ 57 (Nov. 6) [hereinafter Oil Platforms].

[94] *But cf.* Caren Myers Morrison, *Body Camera Obscura: The Semiotics of Police Video*, 54 AM. CRIM. L. REV. 791 (2017) (critiquing notion that real-time recordings of violent interactions can provide objective evidence of what occurred).

[95] John Markoff, *Fearing Bombs That Can Pick Whom to Kill*, N.Y. TIMES (Nov. 11, 2014), https://www.nytimes.com/2014/11/12/science/weapons-directed-by-robots-not-humans-raise-ethical-questions.html.

[96] DINSTEIN, *supra* note 23, at 203.

that its use of force was in self-defense.[97] A purely objective standard would ask only whether State Z had actually launched an armed attack, ignoring what State X did, or should, have believed. A purely subjective standard would ask only whether State X believed an attack was underway, or perhaps would defer entirely to whatever State X says its subjective belief was.[98] A mixed approach might ask whether Actor State X's actual subjective belief that an attack was underway was objectively reasonable.[99] Generally speaking, the standard for adjudicating self-defense claims is predominantly objective. As the ICJ held in *Oil Platforms*, "the requirement of international law that measures taken avowedly in self-defence must have been necessary for that purpose is strict and objective, leaving no room for any 'measure of discretion.'"[100] However, some scholars have argued that various elements of the self-defense inquiry—including the threshold question of whether an "armed attack" has occurred—leave some room for subjectivity, at least for now.[101]

The introduction of AWS into crisis situations should and likely will hasten the move towards objectivity in the adjudication of self-defense claims. The reason is simple: the difficulties intrinsic to subjective analysis of artificially intelligent decision-making make an objective standard far more attractive. Commentators have explored in detail the problem of the "black box" in artificial intelligence, the notion that "[i]t may be impossible to tell how an AI that has internalized massive amounts of data is making its decisions."[102] Consider the black box problem in the context of two hypothetical post-hoc evaluations of an anticipatory strike, one involving a human soldier and the other involving an AWS. The human soldier can be interviewed to determine if she shot first because she thought she saw an enemy aiming a gun at her, or if her gun discharged accidentally, or if she received faulty Rules of Engagement, and so

[97] Oil Platforms, *supra* note 93, at ¶ 57.

[98] *See generally* David Kaye, *supra* note 82, at 149-52 (discussing subjectivity and objectivity in self-defense claims).

[99] For a discussion of the doctrinal history of the tension between objectivity and subjectivity in determining state responsibility, *see* David K. Linnan, *Iran Air Flight 655 and Beyond: Free Passage, Mistaken Self-Defense, and State Responsibility*, 16 YALE J. INT'L L. 245, 354-66 (1991).

[100] Oil Platforms, *supra* note 93, at 196. *See also* Case Concerning Military and Paramilitary Activities Against Nicaragua, Judgment, 1986 I.C.J. 14, ¶282 (June 27) ("[W]hether a measure is necessary to protect the essential security interests of a party is not . . . purely a question for the subjective judgment of the party: the text does not refer to what the party 'considers necessary' for that purpose.") [hereinafter Nicaragua].

[101] David Kaye, *supra* note 82, at 149-52; Norman G. Printer, *The Use of Force Against Non-State Actors Under International Law: An Analysis of the U.S. Predator Strike in Yemen*, 8 UCLA J. INT'L L. & FOREIGN AFF. 331, 338-39 (2003) ("[T]he determination of whether an aggressive use of force crosses the threshold [contemplated by Article 51 of the U.N. Charter] and triggers the exercise of self-help is a subjective one to be made by the attacked state. Nonetheless, the determination is ultimately subject to legal scrutiny by the international community in conformity with the preceding standard.").

[102] Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J. L. & TECH. 889, 891 (2018). *See also* Davide Castelvecchi, *Can We Open the Black Box of AI?*, NATURE (Oct. 5, 2016), https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731 (describing the black box problem).

on. But humans have not figured out how to get a sophisticated AI to tell us precisely how and why it made a given decision in a way that is both comprehensible and comprehensive.[103] Given the difficulty, bordering on impossibility, of asking an AWS about its subjective mindset at the time it launched an anticipatory strike, an objective standard provides the only manageable mode of inquiry.

This prediction, if correct, suggests more accountability for violations of international law. First, shifting to a purely objective standard would simplify adjudication. Rather than delving into actors' mindsets amidst the fog of war, an adjudicator will need only analyze whether the threat of armed attack was *actually* so "instant" and "overwhelming" that it left "no choice of means, and no moment for deliberation."[104] This simplification means that an actor's perceptions of a threat are no longer a justification for its actions; concomitantly, one would expect that a lower proportion of actors claiming self-defense could do so effectively.

Second, because the party claiming self-defense bears the burden of establishing the factual basis for that claim, a purely objective standard would incentivize that party to design and operate AWS to maximize the amount of data recorded about the event, *i.e.*, to create a 'paper trail' to show the lawfulness of their actions. More information about an alleged use of self-defense, in turn, provides the opportunity for more transparency. An analogy familiar to anyone with a basic understanding of how consumer laptops work may be helpful: RAM (random access memory) is to an AWS' data analysis as a hard drive is to an AWS' collection and preservation of its own thought processes and observational data. My argument is that the de-emphasis on what happens in an AWS' RAM and the increased emphasis on what is stored in its hard drive will make it easier for investigators and adjudicators to discern exactly what happened, and thus who (if anyone) violated what laws.[105]

One exception to the foregoing may be found in the post-hoc analysis of human decisions made with the help, or at the recommendation, of artificially intelligent battle-management systems. As Ashley Deeks, Noam Lubell, and Daragh Murray point out, autonomous systems will not only alter the battlespace

[103] *See* Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/; Cliff Kuang, *Can A.I. Be Taught to Explain Itself?*, N.Y. TIMES. MAG. (Nov. 21, 2017), https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

[104] *See supra* note 82 and accompanying text.

[105] I use the RAM and hard-drive analogy only to illustrate the distinction between the analysis of data and the collection and storage of data. Of course, the system architecture of an AWS may be far more complex than the traditional RAM/hard-drive distinction.

as weapons platforms, but also as artificially intelligent systems integrated into C4ISR (Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance).[106] This coheres with the U.S. DOD Defense Science Board's systems-level approach, which conceives of autonomy as "the explicit allocation of cognitive functions and responsibilities between the human and computer to achieve specific capabilities."[107] When the unit of analysis is a combined human-AI decisionmaker, some post-hoc subjective inquiry into how and why the relevant humans made decisions with support from an artificially intelligent battle-management system may be appropriate. The contours of such an inquiry, of course, would depend on the situation and command system in question.

C. *AWS and Mistaken Self-Defense*

AWS raise the possibility that computer glitches amid the fog of war will result in tragedy. The *U.S.S. Vincennes* is a cautionary tale. In 1988, the crew of the *Vincennes* made a horrible error, shooting down an Iranian civilian airliner over the Persian Gulf that the crew of the *Vincennes*—based on the reports of the ship's computerized Aegis defense system—believed to be a military aircraft.[108] Other mistakes involving computer-enabled weapons systems have also resulted in "friendly fire" casualties, such as the Patriot incidents described in Section II.A. As humans increasingly rely on computer-generated assessments of whether an armed attack is underway, and especially when time pressures force human soldiers "out of the loop" in determining a response,[109] the probability that an AWS will mistakenly shoot when it should have held its fire will rise dramatically. Stephanie Carvin puts this in terms of Charles Perrow's classic Normal Accident Theory, arguing that AWS are "tightly coupled, highly interactive complex systems" and consequently, "accidents will inevitably occur."[110] In the context of a tense standoff, such a mistake could result in what the DOD's Directive on Autonomy in Weapons Systems euphemistically terms an "unintended engagement."[111]

[106] Deeks et al., *supra* note 16, at 6 ("The proliferation of machine learning might prompt political and military officials to . . . rely more heavily on detailed machine learning risk assessments when deciding whether and how to use force.").

[107] *Task Force Report: The Role of Autonomy in DoD Systems*, OFF. UNDER SECRETARY DEF. ACQUISITION, TECH. & LOGISTICS (July 2012) at 4, https://fas.org/irp/agency/dod/dsb/autonomy.pdf.

[108] Linnan, *supra* note 99, at 246-47.

[109] For a discussion of AWS and Boyd's OODA (Observe, Orient, Decide, Act) "loop," *see generally* Schmitt & Thurnher, *supra* note 36.

[110] Stephanie Carvin, *Normal Autonomous Accidents: What Happens When Killer Robots Fail?* (Mar. 1, 2017) (unpublished manuscript) at 3, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3161446.

[111] DoDD 3000.09, *supra* note 20, at 15 (defining an "unintended engagement" as "[t]he use of force resulting in damage to persons or objects that human operators did not intend to be the targets of U.S. military operations, including unacceptable levels of collateral damage beyond those consistent with the law of war, ROE, and commander's intent.").

Normally, a claim of mistaken but reasonable self-defense is no claim of self-defense at all. As the ICJ held in its *Nicaragua* decision, the use of force in self-defense must be "proportional to the armed attack and necessary to respond to it."[112] A shot fired in self-defense cannot be "necessary to respond to" an armed attack if no armed attack has occurred.[113] As David Linnan persuasively argues, the notion of a "mistaken self-defense" justification for the use of force under international law finds scant precedential or scholarly support.[114] Thus, a State deploying AWS in a crisis can expect to be held responsible if and when a "Normal Accident" occurs in which its autonomous weapon system fires in mistaken self-defense.

Two factors may alter this analysis. First, as I have argued elsewhere, States that wish to avoid both conflict and the audience costs of allowing a perceived transgression to go unpunished may prefer, as an off-ramp, to blame a use of force on a technical glitch.[115] In a recent real-world example, President Trump suggested that Iran had mistakenly shot down a U.S. drone, attributing the use of force to a rogue "loose and stupid" military commander.[116] As a legalistic corollary, States may prefer to resolve such incidents through what Rebecca Crootof has termed "war torts."[117] Professor Crootof argues that international tort law could solve the "accountability gap" that emerges when an AWS commits an atrocity that cannot qualify as a war crime because an autonomous system cannot possess intention or be legally reckless.[118] Under her proposal, a State whose AWS used force in mistaken self-defense would owe compensation but would not need accept moral blameworthiness.[119] Applying Crootof's remedy-based framework to the narrow class of illegal uses of force when an AWS (allegedly) fires in mistaken self-defense may also serve as a sort of "emergency release valve," reducing the domestic audience costs of de-escalation and allowing leaders to avoid further conflict.[120]

[112] Nicaragua, *supra* note 100, ¶176.

[113] An action must be *objectively* necessary for self-defense; an actor's beliefs, reasonable or not, are irrelevant. *See supra* notes 95-105 and accompanying text.

[114] Linnan, *supra* note 99.

[115] Leys, *supra* note 41, at 60-61. The theoretical basis for this claim is James Fearon's model of crisis escalation. At risk of oversimplification, Fearon posits that because "side payments" to resolve a dispute are almost always possible, choosing war is irrational. The answer to the paradox of why wars nevertheless occur is that policymakers view the domestic audience costs of de-escalation as outweighing the expected costs of war. *See* Fearon, *Rationalist Explanations*, *supra* note 34; Fearon, *Domestic Political Audiences*, *supra* note 34.

[116] Demetri Sevastopulo et al., *Trump Says 'Hard to Believe' Downing of U.S. Drone Was Intentional*, FIN. TIMES (June 20, 2019), https://www.ft.com/content/8d41d4a2-930e-11e9-aea1-2b1d33ac3271.html.

[117] Crootof, *War Torts*, *supra* note 38.

[118] *Id.* at 1353.

[119] *Id.* at 1389-94.

[120] Using claims of mistake as an off-ramp may deescalate crises even when the initial strike was *not* a mistake. For example, India and Pakistan have long used "untruths . . . [as] a much-needed off-ramp for dampening tensions" amid border disputes. C. Christine Fair, *India's and Pakistan's Lies Thwarted a*

Second, the liability of a State for mistaken self-defense is intrinsically related to the attributability of the use of force to that State. The doctrine of state responsibility's legal fiction of attribution makes sense when a human soldier makes the mistake; after all, "[s]tates act through human beings."[121] But the fiction that "l'état, c'est un robot" may become untenable as States increasingly rely on AWS. For example, some have raised the possibility of products liability for AWS' violations of international law, implying that the violation is more readily attributable to the manufacturer or designer, rather than the State.[122] Hacking or spoofing may also place an AWS's actions outside any meaningful sense of control or commander's intent.[123] Compounding these problems, militaries employing AWS may wish to keep the scope of authority granted to autonomous systems opaque for obvious strategic reasons—making it all but impossible to determine if an AWS' allegedly *ultra vires* act is incidental to the empowered capacity (in which case the action is attributable to the State) or if the AWS' act falls outside the grant of authority (in which case the action is *not* attributable to the State).[124]

So, if an AWS fires first in a crisis, a state may argue that it did not violate *jus ad bellum*. Not because the AWS (and thus the State) acted in justifiable—but mistaken—self-defense, but because the autonomous system's actions cannot be attributed to the State *at all*. Therefore, the argument goes, the State did not violate the international legal prohibition on the use of force, because the State *qua* State did not use force. Whether this view is correct or not, the point is that it would be in a State's interest to so argue. This creates a tension with the widely held view under *jus in bello* of commander responsibility for their AWS.[125] If States can advance their interests and avoid dangerous escalation by arguing that they are not responsible for the alleged errors of their

War—For Now, THE ATLANTIC (Mar. 8, 2019), https://www.theatlantic.com/international/archive/2019/03/india-pakistan-kargil-kashmir/584392/. By misinforming their respective electorates, India and Pakistan reduced the domestic audience costs of de-escalation, potentially averting a war between nuclear-armed states. *See* Fearon, *Domestic Political Audiences*, *supra* note 34.

[121] Linnan, *supra* note 99, at 364.

[122] PATRICK LIN ET AL., AUTONOMOUS MILITARY ROBOTICS: RISK, ETHICS, AND DESIGN 56-57 (2008), https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1001&context=phil_fac.pdf. *But see* Daniel N. Hammond, *Autonomous Weapons and the Problem of State Accountability*, 15 CHI. J. INT'L L. 652, 665-67 (2015) (criticizing calls for programmer or products liability); *Losing Humanity: The Case Against Killer Robots*, *supra* note 35, at 43-44 (same). *See generally* Tim McFarland & Tim McCormack, *Mind the Gap: Can Developers of Autonomous Weapon Systems be Liable for War Crimes?*, 90 INT'L L. STUD. 361 (2014).

[123] *See, e.g.*, Scharre, *Operational Risk*, *supra* note 8, at 38-39 ("Adversaries would have strong incentives to hack such [autonomous] systems, either directly through malware or via behavioral hacking . . . .").

[124] TALLINN MANUAL, *supra* note 24, at 90-91 (addressing this concept in the context of cyberattacks).

[125] *See, e.g.*, Charles J. Dunlap, Jr., *Accountability and Autonomous Weapons: Much Ado About Nothing*, 30 TEMPLE INT'L & COMP. L.J. 63, 70 (2016) ("In truth, it is not complicated to find command accountability for directing the use of any weapon without a reasonable belief that doing so would comply with the law of war.").

AWS under *jus ad bellum*, then what becomes of state responsibility for atrocities committed by AWS under *jus in bello*?

D. *Self-Defense of What? AWS and* Jus ad Bellum *Proportionality*

AWS may throw a wrench into conventional evaluations of *jus ad bellum* proportionality (a distinct concept from its *jus in bello* counterpart). Under established *jus ad bellum* principles, force used in self-defense must be proportionate to the force used in the armed attack giving rise to the right of self-defense.[126] Heather Roff, in one of the few articles to address AWS through a *jus ad bellum* lens, argues that a State employing AWS in a defensive posture can never satisfy this proportionality requirement.[127] She argues that because AWS cannot die, the State deploying them will be less likely to terminate hostilities than if it relied on human soldiers.[128] The primary legal weakness of Roff's analysis is that it relies on a philosophical—rather than legalistic—approach to *jus ad bellum* proportionality. Her framework "weigh[s] only the relevant goods of waging war," that is, preventing the completion or recurrence of a single, aggressive use of force, "against *all* of the foreseen evils of doing so," including every foreseeable harm resulting from escalation.[129] This conception of proportionality is out-of-step with accepted legal understandings of *jus ad bellum* proportionality, which is confined to comparing "the quantum of [aggressive] force and [defensive] counter-force used, as well as the casualties and damage sustained."[130] "There is no support in the practice of States for the notion that [*jus ad bellum*] proportionality remains relevant—and has to be constantly assessed—throughout the hostilities in the course of war."[131] Thus, bracketing whether Roff is correct from a philosophical point of view, her argument is not legally persuasive.

Despite the legal weaknesses in Roff's analysis, she does make one point that illuminates an interesting problem of AWS and *jus ad bellum* proportionality. Roff argues that if one confines the proportionality analysis to a given engagement between AWS not involving humans—say, two autonomous UAVs dogfighting far from any human-occupied aircraft—"one would have to

[126] *See, e.g.*, Case Concerning Armed Activities on the Territory of the Congo (D.R.C. v. Uga.), Judgment, 2005 I.C.J. 168, ¶147 ("[T]he taking of airports and towns many hundreds of kilometres from Uganda's border would not seem proportionate to the series of transborder attacks it claimed had given rise to the right of self-defence, nor to be necessary to that end.").

[127] Roff, *Lethal Autonomous Weapons*, *supra* note 16, at 37.

[128] *Id.* at 37-38.

[129] *Id.* at 41.

[130] DINSTEIN, *supra* note 23, at 262.

[131] *Id.*

rely on arguments concerning property damage."[132] The framing of AWS as property—or, in the terms of the laws of war, materiel—raises the question of whether an attack on an autonomous weapon renders lawful the same defensive use of force as would a similar attack on a human.[133] After all, if States deploy AWS in part because robots are not shipped home in flag-draped coffins,[134] the implication is that States (rightly) view the destruction of AWS as less serious than the killing of human soldiers. Therefore, if an AWS detects an incoming attack that threatens itself but not any nearby humans, and determines that returning fire would endanger enemy humans, the principle of *jus ad bellum* proportionality as currently understood may require the autonomous system to hold its fire.

Not all experts may agree on this point. The Tallinn Manual 2.0, for example, assumes without explanation that an attack on "*property* or persons" may qualify as an "armed attack" within the meaning of Article 51.[135] The logic appears to be based on the Law of Armed Conflict's well-established rule that military materiel is a legitimate target during armed conflicts.[136] But that rule falls under the rubric of *jus in bello*, and may not mechanically translate to the domain of *jus ad bellum* where the rules of proportionality are generally stricter.[137] I do not expect the question of whether an attack on non-human property (i.e., an AWS) can qualify as an armed attack to be settled by this Note. I intend rather to make the point that already-difficult questions of

[132] Roff, *Lethal Autonomous Weapons*, *supra* note 16, at 45. Roff goes on to argue that because "common law and US jurisprudence prohibit using lethal force in defense of property," damage to AWS cannot satisfy the proportionality requirement of *jus ad bellum*. *Id.* This argument fails to recognize that the laws of armed conflict are fundamentally different from domestic criminal law. For example, the laws of war permit killing by mistake (assuming the killer tried their best to minimize collateral damage). Domestic law does not. More obviously, U.S. criminal law would not govern a military conflict between, say, Luxembourg and Micronesia.

[133] Both enemy materiel and enemy combatants are considered lawful targets under *jus in bello*. *See* Schmitt & Thurnher, *supra* note 36, at 246-47 (2013) (discussing AWS' ability to distinguish between lawful targets (combatants and military objects) and unlawful targets (noncombatants and civilian objects)).

[134] For a discussion of the political dynamics of states' decisions to deploy AWS, *see, e.g.*, Frank Sauer & Niklas Schörnig, *Killer Drones: The 'Silver Bullet' of Democratic Warfare?*, 43 SEC. DIALOGUE 363 (2012).

[135] TALLINN MANUAL, *supra* note 124, at 346 (emphasis added). *But see id.* at 340 (arguing that non-kinetic attacks may constitute armed attacks "because the ensuing consequences can include serious suffering or death," which may imply that an attack which could *not* cause bodily harm would not trigger a right to self-defense under Article 51).

[136] *See* Yoram Dinstein, *Legitimate Military Objectives Under Current Jus in Bello*, 78 INT'L L. STUDIES 139, 146-47 (2002).

[137] *See, e.g.*, Robert D. Sloane, *The Cost of Conflation: Preserving the Dualism of* Jus ad Bellum *and* Jus in Bello *in the Contemporary Law of War*, 34 YALE J. INT'L L. 47, 52-53 (2009) ("Briefly, *ad bellum* proportionality asks whether the initial resort to force or particular quantum of force used is proportional to the asserted *casus belli*. . . . *In bello* proportionality tries to limit needless suffering in war regardless of the *ad bellum* legitimacy . . . . It asks whether each *particular* strike will cause civilian harm that 'would be excessive in relation to the concrete and direct military advantage anticipated.'" (citation omitted)).

proportionality become even more important in the strategic and domestic political context of autonomous weapons.

E. *AWS and "Armed Attack"*

The previous section dealt with attacks *on* AWS; this section deals with the mirrored question of attacks *by* AWS. If the AWS of State X use force against State Z without the proximate involvement of human commanders of State X, then under existing ICJ precedent State Z may not have any legal right to self-defense, anticipatory or otherwise.[138] This is because the right to self-defense under the UN Charter arises only "if an armed attack occurs."[139] Under ICJ precedent, it is not entirely clear that sending non-human soldiers to inflict harm on an adversary constitutes "an armed attack." In the *Nicaragua* case, the ICJ held that "an armed attack . . . includ[es] not merely action by regular armed forces across an international border, but also 'the sending by or on behalf of a State of armed bands, groups, irregulars, or mercenaries.'"[140] In other words, attacks *by humans* acting on behalf of a State count, whether or not they are incorporated into the aggressive State's formal military command structure. But "while the concept of an armed attack includes the despatch by one State of armed bands into the territory of another State, the supply of arms and other support to such bands cannot be equated with armed attack."[141] Per *Nicaragua*, "the provision of weapons or logistical or other support" does not rise to the level of an "armed attack" within the meaning of Article 51.[142] The *Nicaragua* judgment's juxtaposition of sending human fighters on the one hand, and sending weapons or military materiel on the other, raises the question of whether AWS count as soldiers/armed bands or weapons.[143] This question is easily disposed of when the offensive use of an AWS is the result of a human operator's command: if a soldier of State X tells an AWS to shoot down planes of State Z, it would be

[138] Note that this particular scenario does not necessarily involve an autonomous weapon using force in anticipatory self-defense; the introduction of AWS changes questions of self-defense on both sides of the equation.

[139] U.N. Charter, art. 51.

[140] Nicaragua, *supra* note 100, ¶195 (citation omitted).

[141] *Id.* at ¶247. *Cf.* Michael J. Glennon, *The Fog of Law: Self-Defense, Inherence, and Incoherence in Article 51 of the United Nations Charter*, 25 HARV. J. L. & PUB. POL'Y 539, 541 (2002) ("Providing weapons and logistical support to terrorists does not constitute an 'armed attack'.").

[142] Nicaragua, *supra* note 100, ¶195.

[143] The ICJ's holding in *Nicaragua* could be read to mean that the provision of weapons cannot give rise to a right to self-defense because it is not sufficiently *attributable* to the allegedly belligerent state. This has been referred to as the "'*rationae personae*' aspect [of *Nicaragua*] – that is, from whom the attack emanates." Abdulqawi A. Yusuf, *The Notion of 'Armed Attack' in the* Nicaragua *Judgment and Its Influence on Subsequent Case Law*, 25 LEIDEN J. INT'L L. 461, 462 (2012). But *Nicaragua*'s definition of "armed attack" also encompasses what Judge Yusuf refers to as "the *rationae materiae* aspect of the concept," defining the level of force which constitutes an armed attack. *Id.* at 461. It is this latter notion which is applicable here.

absurd to suggest that the delegation of targeting and firing functions renders such an action not an "armed attack."

But the question is harder when the use of force in question is not the direct result of human intention. The ICJ has suggested repeatedly that naval mines, for example, do not rise to the level of armed attack.[144] Obviously, there are important distinctions between a naval mine and a truly autonomous weapon system. The ICJ, however, has not drawn a clear distinction based on the degree of autonomy possessed by the attacking unit; rather, its distinction appears to rest on whether the attacking State sends humans or weapons to cause damage.

For purposes of the *jus ad bellum* inquiry, international courts and policymakers could certainly reverse course and assimilate AWS into the former category, such that their offensive use could constitute an "armed attack." But this would seem incoherent with international efforts to ban or regulate AWS *as weapons*, in the vein of blinding lasers, landmines, or chemical weapons.[145] If AWS are "weapons" for the purposes of International Humanitarian Law (which falls under the *jus in bello* umbrella), then shouldn't they also be "weapons" for the purposes of *jus ad bellum*? Thus, in the case of an offensive action taken by State X's AWS not at the direct command of State X's human commander, it is not clear that State Z has any right to use force in self-defense.

The assertion that damage caused by an AWS cannot amount to an armed attack giving rise to a right of self-defense may cause some justifiable skepticism. And, of course, the ICJ or other adjudicators may not apply the *Nicaragua* distinction between soldiers and materiel so rigidly. There are at least two paths to square this particular circle. First, an adjudicating body might decide

[144] Oil Platforms, *supra* note 93, ¶72 ("The question is therefore whether [the mining of the U.S.S. *Samuel B. Roberts*] sufficed in itself to justify action in self-defence, as amounting to an 'armed attack'. The Court does not exclude the possibility that the mining of a single military vessel might be sufficient to bring into play the 'inherent right of self-defence'; but in view of all the circumstances . . . the Court is unable to hold that the [U.S. uses of force] have been shown to have been justifiably made in response to an 'armed attack on the United States by Iran, in the form of the mining of the USS *Samuel B. Roberts*."); The Corfu Channel Case (U.K. v. Alb.), Judgment, 1949 I.C.J. 4, 35 ("The United Kingdom . . . has classified 'Operation Retail' [entering Albanian waters for the purposes of clearing mines] among methods of self-protection or self-help [following the destruction of two UK ships]. The Court cannot accept this defense . . . .").

[145] *See, e.g.*, John Lewis, *The Case for Regulating Fully Autonomous Weapons*, 124 YALE L.J. 1309 (2015) (arguing that international regulation of landmines provides a model for regulating AWS); Crootof, *Killer Robots*, *supra* note 8, at 1883-90 (evaluating calls to ban AWS through the lens of other weapon prohibition efforts). Rebecca Crootof has argued persuasively that whether policymakers and adjudicators classify AWS as weapons, combatants, or something in between will determine the answer to many important *jus in bello* questions; my argument is simply that the same may be true in the *jus ad bellum* context of determining whether an "armed attack" has occurred. Rebecca Crootof, *Autonomous Weapon Systems and the Limits of Analogy*, 9 HARV. NAT'L SEC. J. 51, 55 (2018) ("[T]he weapon and combatant analogies for Autonomous Weapon Systems are at odds with each other, insofar as they implicate distinct regulatory regimes. . . . Given this distinction, selecting the weapon or combatant analogy will predetermine the answers to many troubling legal questions . . . .").

to revisit *Nicaragua*'s admittedly ambiguous language, recasting its soldier/materiel distinction as really being about the level of damage rather than the type of incursion. Second, an adjudicating body may follow the example of the Claims Commission in *Ethiopia v. Eritrea*. There, the Commission held that "[l]ocalized border encounters between small infantry units, even those involving the loss of life, do not constitute an armed attack for purposes of the Charter."[146] Framing at least low-level attacks by AWS without proximate human involvement as something akin to "localized border encounters," giving rise to state responsibility but not a right to self-defense, is a plausible and potentially escalation-dampening legal rule.[147]

IV. NEW "RULES OF THE ROAD": HOW THE LAW SHOULD TREAT AWS IN CRISIS SITUATIONS

The introduction of AWS into crisis situations will present new legal and strategic challenges. The prior sections sought to illuminate those difficulties; we now turn to what new rules might address these issues.

I do not here deal with enforcement in great detail. The question of how to make sure these rules are followed is of course of great importance, but analytically separate from this analysis. One would hope that because escalation to conflict is a uniquely costly way of resolving disputes, States would adopt rules that reduce the risk of escalation out of pure self-interest. I view this as the best possible enforcement mechanism, because if a State truly believes that deployment and use of AWS in violation of the following rules is in its own interest, there is only so much weight that outrage from the International Committee of the Red Cross and condemnation from a weapons review panel will carry.[148]

To reduce the twin risks of escalation and violations of international law, militaries should consider imposing a set of operational and programmatic constraints on AWS in crisis situations. The first of these safeguards would require AWS in swarming configurations to absorb the first blow before retaliating, unless doing so would endanger human life. The second and third relate to AWS' strategic awareness: a "*jus ad bellum* switch" and an "auto-off" feature. The fourth proposes an "orange box" rule to ensure the preservation of data regarding the decision to strike. These rules would improve compliance with

[146] Eri.-Eth. Claims Comm'n, Partial Award: *Jus Ad Bellum* – Ethiopia's Claims 1-8, XXVI R.I.A.A. 457, ¶11 (Dec. 19, 2005).

[147] *But see* Christine Gray, *The Eritrea/Ethiopia Claims Commission Oversteps Its Boundaries: A Partial Award?*, 17 EUR. J. INT'L L. 699 (2006) (criticizing the above-referenced decision).

[148] *See* Schmitt & Thurnher, *supra* note 36, at 271-76 (2013) (discussing weapons review obligations under international law with respect to AWS).

international law, dampen escalation, reduce the chance of illegal wars, and save lives.

A. *Swarming AWS and First-Blow Absorption*

In some circumstances, swarming Autonomous Weapon Systems should be required to absorb the first blow before firing back. AWS will enter the battlefield in networked swarms.[149] For example, the U.S. military has tested a swarm of around one hundred small winged UAVs called "Perdix" drones, which are released from the back of a fighter jet and then coordinate with each other to accomplish their objective.[150] In addition to improving AWS' lethality and offensive capabilities, "swarm resiliency" will replace "[p]latform survivability."[151] That is, the ability of a swarm of AWS to lose a fraction of its members and continue functioning as a collective will become more important than the ability of a single unmanned system to survive an attack. As Paul Scharre notes, "[i]ndividual platforms need not be survivable if there are sufficient numbers of them such that the whole is resilient against attack."[152] He continues, swarming "allows the *graceful degradation* of combat power as individual platforms are attrited, as opposed to a sharp loss in combat power if a single, more exquisite platform is lost."[153] In other words, for an AWS swarm, the cost of absorbing the first blow may be quite low. So too is the moral (and political) cost, relative to an attack on a group of human soldiers—an AWS cannot, in Thomas Schelling's words, "die heroically, dramatically, and in a manner that guarantees that the action cannot stop there."[154]

Swarming AWS should be programmed not to fire in anticipatory self-defense unless it is necessary to protect human life. Imagine a swarm of autonomous UAVs patrolling a disputed border. The swarm discerns that the radar of a surface-to-air missile has locked on to them and calculates that an incoming missile will destroy between 5-10 percent of its UAVs. Under the proposed rule, the UAVs may take evasive action, but may not fire on the missile or the launch site until the missile has actually detonated. Or imagine that the UAVs detect a child pointing what appears to be an RPG at them. One would

[149] Chris Jenks, *The Gathering Swarm: The Path to Increasingly Autonomous Weapon Systems*, 57 JURIMETRICS 341, 351-55 (2017).

[150] Jamie Condliffe, *A 100-Drone Swarm, Dropped from Jets, Plans Its Own Moves*, MIT TECH. REV. (Jan. 10, 2017), https://www.technologyreview.com/s/603337/a-100-drone-swarm-dropped-from-jets-plans-its-own-moves.html.

[151] Paul Scharre, *Robotics on the Battlefield Part II: The Coming Swarm*, CTR. NEW AM. SEC. (2014), https://s3.amazonaws.com/files.cnas.org/documents/CNAS_TheComingSwarm_Scharre.pdf?mtime=20160906082059.pdf.

[152] *Id.*

[153] *Id.*

[154] THOMAS SCHELLING, ARMS AND INFLUENCE 47 (1966).

not expect a human pilot to wait to be fired on in the first scenario, and the second presents a horrifying ethical conundrum. But for a swarm of AWS, neither situation presents a choice between taking a life or risking your own.

Forcing a swarm of AWS to absorb the first blow may marginally degrade combat ability, but the strategic and legal benefits of such a rule are clear and dramatically outweigh the costs. Strategically, requiring State X to suffer actual losses before returning fire reduces the escalatory possibility of an "unintended engagement" in which State X fires the first shot.[155] Legally, such a rule would ameliorate the evidentiary issues addressed in Section III.A, because evidence of an actual attack on a swarm of AWS would help circumvent the "black box" issue.[156] It would also reduce the risk of an AWS firing in mistaken anticipatory self-defense, as considered in Section III.C. And it would simplify the proportionality analysis discussed in Section III.D by allowing an AWS to calibrate its response based on an actual battle-damage assessment, rather than a probabilistic projection of an incoming attack's potential impact.[157] Finally, the period between the detection of an incoming attack and the completion of that attack will allow more time for "meaningful human control," which may alleviate certain *jus in bello* concerns beyond the scope of this Note.[158]

The proposed rule includes an important exception: AWS detecting an incoming attack should be allowed to fire in anticipatory self-defense when the attack threatens human life. As argued in Section III.D, if an attack threatens both AWS and humans (as opposed to merely AWS), then the principle of proportionality would permit greater force to be used. The same goes for the necessity principle. Assuming that the attacker presents an ongoing threat to human life (e.g., an artillery battery preparing to fire on a populated area), anticipatory or "interceptive"[159] self-defense may be necessary to protect those lives, whereas a threat only against AWS may not necessitate the same level of counterforce. This exception, and the accompanying proportionality and necessity analyses, follow directly from the observation that robots are worth

[155] DoDD 3000.09, *supra* note 20, at 15.
[156] *See* Bathaee, *supra* note 102.
[157] *See* DINSTEIN, *supra* note 23, at 262 (arguing that the *jus ad bellum* proportionality analysis requires comparing "the quantum of force and counter-force used, *as well as the casualties and damage sustained.*" (emphasis added)); Deeks et al., *supra* note 16, at 10 (discussing the use of autonomous systems to inform proportionality and necessity analyses).
[158] *See generally* Michael C. Horowitz & Paul Scharre, *Meaningful Human Control in Weapon Systems: A Primer*, CTR. NEW AM. SEC. (2015), https://s3.amazonaws.com/files.cnas.org/documents/Ethical_Autonomy_Working_Paper_031315.pdf?mtime=20160906082316; Rebecca Crootof, *A Meaningful Floor for "Meaningful Human Control,"* 30 TEMPLE INT'L & COMP. L.J. 53 (2016).
[159] DINSTEIN, *supra* note 23, at 204-05.

less than human life. AWS will not make war "bloodless."[160] But conditioning their actions on the probability of bloodletting—on both sides—may make war marginally more humane and less likely to spiral out of control.

B. *AWS and Command-and-Control*

An AWS cut off from its C2 network is more likely to use force illegally or to fail to use force when it would be legal and strategically desirable to do so. Notably, as discussed in Section II.D, an AWS cut off from its human commanders might fire in an illegal and escalatory fashion if it were under the mistaken belief that hostilities were ongoing. [161] Two safeguards may ameliorate this risk. First, States developing AWS should endeavor to establish a "*jus ad bellum* switch" in AWS, triggered by resilient communications systems. Second, States may consider installing a time-delayed "auto-off" function in situations where C2 is completely degraded.

First, to prevent AWS using force under the mistaken assumption that hostilities are ongoing, militaries developing AWS should include a "*jus ad bellum* switch," which can be triggered by highly resilient communications systems. As Paul Scharre notes, "[c]ommunications in contested areas is not an all-or-nothing proposition."[162] Militaries may be able to maintain minimal contact with AWS in denied environments, insufficient to directly control[163] AWS but potentially enough to let an AWS know whether a broader state of hostilities exists.[164] Such a capability would minimize the risk that an AWS mistakenly uses force in violation of international law. This system would also function as a two-way street; an AWS in a communications-limited environment could also let its broader command structure know whether an adversary has launched an "armed attack," triggering a right to self-defense.

Imagine that amid a conventional conflict in the South China Sea, State X's satellites have been destroyed. State X relies heavily on its satellites to control its unmanned submersibles, which have an autonomous feature. But using a backup communication system of Very Low Frequency (VLF) radio transmissions, which have extremely limited bandwidth, State X can maintain a rudimentary uplink with its unmanned subs scattered across the Pacific. State X

[160] SCHARRE, ARMY OF NONE, *supra* note 7, at 303 (2018).
[161] This is a *jus ad bellum* spin on Roff's "Strategic Robot Problem." *See* Roff, *Lethal Autonomous Weapons*, *supra* note 16, at 211.
[162] Paul Scharre, *Centaur Warfighting: The False Choice of Humans vs. Automation*, 30 TEMPLE INT'L & COMP. L.J., 151, 161-64 (2016) [hereinafter Scharre, *Centaur Warfighting*].
[163] Taking direct command of an AWS—that is, deactivating its autonomous mode—would likely require substantially more bandwidth than allowing it to continue operating autonomously or ordering it to hold its fire. For more on the AWS-bandwidth connection, see *supra* note 71 and accompanying text.
[164] Scharre, *Centaur Warfighting*, *supra* note 162.

could use this system to inform its subs whether a state of conflict is ongoing, in which case the subs' pre-existing programming for compliance with *jus in bello* controls, or whether the conflict has ceased, in which case the sub knows that it can only fire in line with the *Caroline* standard. Concretely, an autonomous sub that stumbles upon State Y's destroyer under the former condition could launch a torpedo; but under the latter scenario, the submarine could only fire if the destroyer posed a sufficiently immediate threat.

Second, in completely communications-denied environments, AWS may need an "auto-off" switch. Such a feature—analogous to the timed deactivation mechanisms on coffee pots or so-called "self-deactivating" mines[165]—would switch an AWS from an aggressive to a defensive posture after a set amount of time. Take the South China Sea hypothetical posed above, but now assume that each side's capacity to maintain communication with its respective AWS has been completely eliminated. Before the destruction of its command-and-control systems, State X activates an autonomous submersible in a shipping lane to hunt and destroy State Z's warships. Once a preset time—say, 72 hours—passes without a periodic confirmation from the submersible's commanding military that hostilities are ongoing, the submersible shifts back into its pre-hostilities setting: on alert, but set not to fire except when permissible under the *Caroline* standard. The submersible, in this setting, would not pursue targets of opportunity. This feature would not guarantee total compliance with the laws of war. But it would significantly minimize the risk of a "Battle of New Orleans"[166] or "Japanese holdout"[167] violation.

Of course, both these approaches raise the risk that State X's AWS might be under-aggressive in a way that redounds to the strategic benefit of State Z. But in neither instance is the risk unreasonable. First, the "*jus ad bellum* switch" might be hacked.[168] The task of preventing that from occurring is a subset of the challenges of maintaining C2 integrity in contested environments; and in any case, a military unable to ensure basic communications security is unlikely to prevail in a conflict. Even for AWS that operate stealthily or behind enemy lines, mitigating the hacking danger is conceptually different in degree,

[165] *Smart Weapons: Kill Switches and Safety Catches*, ECONOMIST (Nov. 30, 2013), https://www.economist.com/technology-quarterly/2013/11/30/kill-switches-and-safety-catches.html, (discussing self-deactivating mines). *See also* DoDD 3000.09, *supra* note 20, at 2 ("[AWS shall] . . . [c]omplete engagements in a timeframe consistent with commander and operator intentions and, if unable to do so, terminate engagements or seek additional human operator input before continuing the engagement.").

[166] *See supra* notes 66-80 and accompanying text.

[167] Leys, *supra* note 41, at 58, 66.

[168] *See, e.g.*, Michal Klincewicz, *Autonomous Weapon Systems, the Frame Problem and Computer Security*, 14 J. MIL. ETHICS 162 (2015) (arguing that the complexity of AWS makes hack-proofing impossible).

but not in kind, from the difficulty of transmitting nuclear launch codes to submarines. Second, the auto-off function might cause an AWS to hold its fire when using force would be legal and more strategically advantageous. But failing to include this feature presents the reverse risk: that an AWS may fire when it would be more advantageous to refrain from using force. The latter danger may violate the laws of war; the former would not. The latter may prove unnecessarily escalatory; the former would promote de-escalation. Moreover, the strategic risks of an AWS entering "auto-off" mode can be mitigated by pre-installing rules of engagement that, like the absorption rule proposed in Section IV.A, allow the system to effectively defend itself.

C. *The "Orange Box" Rule*

An AWS that decides to fire in anticipatory self-defense should be designed to preserve all data regarding its decision before firing. This rule would reduce the chances of hacking or spoofing leading to an unintended conflict. I refer to this as the "orange box" rule, named for the color and shape of the flight recorders installed on aircraft.[169] Recorders have proved crucial in understanding how AI decisions have gone disastrously wrong in the past, most prominently in analyzing plane crashes involving a malfunctioning autopilot.[170] Unlike commercial airliners, however, most AWS will function in highly networked swarms or will be connected to military command-and-control systems. For that reason, it should be relatively straightforward to transmit and back-up this data in close to real-time, rather than storing it onboard an AWS.

To illustrate the "orange box" rule's protection against spoofing, take the Harpy example presented above.[171] Under this rule, if a patrolling Harpy detects an enemy radar and decides that it must strike to prevent a missile launch against a civilian population, it should first transmit as much data as possible to a friendly system for storage and analysis. An adjudicator after the fact would then have access to the AWS' observations and an audit of its internal decision-making process.[172] This information, in turn, would simplify the task of

[169] Although these recorders are frequently referred to as a "black box," this nickname is both visually inaccurate and confuses the concept of the recorder, which is designed to enhance transparency, with the concept of an impenetrable decision-making process, which by its nature is opaque.

[170] *See, e.g.*, Paul Schemm, *Ethiopian Official: Black Box Data Show 'Clear Similarities' Between Ethiopian Airlines, Lion Air Crashes*, WASH. POST (Mar. 17, 2019), https://www.washingtonpost.com/world/ethiopian-official-black-box-data-shows-clear-similarity-between-ethiopian-airlines-lion-air-crashes/2019/03/17/92573222-48d4-11e9-8cfc-2c5d0999c21e_story.html.

[171] *See supra* notes 9-12, 95-96 and accompanying text.

[172] Some readers may worry that I am too cavalier in assuming the "explainability" of a given AI's decision-making process. There is a certain "assume a can opener" quality to proposals that turn on the ability of AI to make itself comprehensible to humans. However, explainability (or "xAI") is a priority

determining whether a given strike objectively met the *Caroline* standard. In the aftermath of the Harpy striking a civilian building, for instance, the adjudicator could parse through the "orange box" data to determine if Hamas had placed a targeting radar on top of the structure.

The "orange box" rule would also mitigate the risk of hacking, in at least three ways. First, by preserving and transmitting the data received about a hacking attempt, the rule would reduce the ability of an adversary to use a zero-day exploit[173] against multiple AWS. With more information regarding a vulnerability, autonomous cyberdefenses will be better positioned to devise a response.[174] Second, even if an adversary manages to hack State X's AWS so that it transmits data falsely indicating that it is still under State X's control,[175] other nearby AWS could identify and destroy a compromised AWS by flagging a difference between their observations of the hacked system and its transmitted data. Third, the preservation and transmittal of data would assist a State whose AWS were hacked to prove that it was not in control of an AWS that fired first, and therefore did not violate *jus ad bellum* by starting a war.

V. CONCLUSION

Despite the risk of AWS starting a war, the dearth of literature regarding autonomy, international crises, and *jus ad bellum* suggests that lawyers and policymakers have not paid nearly enough attention to these questions. This Note has argued that AWS may use force in anticipatory self-defense in at least some circumstances.

for the U.S. military. *See* Sara Castellanos & Steven Norton, *Inside DARPA's Push to Make Artificial Intelligence Explain Itself*, WALL ST. J. (Aug. 10, 2017), https://blogs.wsj.com/cio/2017/08/10/inside-darpas-push-to-make-artificial-intelligence-explain-itself. *See generally* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829 (2019) (discussing the legal implications of xAI). The European Union's General Data Privacy Regulation (GDPR) has also raised the issue's profile by creating a "right to explanation." *See generally* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERK. TECH. L.J. 189 (2019). For the purposes of auditing AWS' decisions to strike in anticipatory self-defense, I assume only that the data stored in the proverbial "orange box" regarding the decision-making process will be of at least some use to an adjudicator, though a perfect understanding of the decision to strike may not be possible.

[173] A "zero-day exploit" refers to a cyber-vulnerability which has never before been used. The name refers to the amount of time which those defending the threatened system have to respond to the attack.

[174] *See generally* Kenneth Anderson, *Why the Hurry to Regulate Autonomous Weapon Systems—But Not Cyber-Weapons?*, 30 TEMP. INT'L & COMP. L.J. 17 (2016) (discussing autonomous cyberweapons).

[175] This tactic was used by the United States in its hacking operation against Iran's nuclear centrifuges. Stuxnet, the worm that infiltrated Iran's centrifugal control systems, caused the centrifuges to spin out of control, while transmitting false data suggesting that the facility was operating normally. *See* David Sanger, *Obama Order Sped Up Wave of Cyberattacks Against Iran*, N.Y. TIMES (June 1, 2012), https://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html.

Furthermore, with the right rules in place, introducing AWS into crises may improve compliance with *jus ad bellum*, reduce escalation, and save lives. Militaries developing AWS should consider installing several technical safeguards into their autonomous systems. First, AWS in swarming configurations should be programmed to absorb the first blow in crisis situations, except when doing so would present a threat to human life. Second, militaries should endeavor to install a "*jus ad bellum* switch" in their AWS and corresponding C2 structures. This system would ideally allow a military to inform its AWS whether or not a state of hostilities exists, even in degraded C2 environments. Third, AWS should have an "auto-off" function. If, during a period of hostilities, the AWS loses contact with its command structure for more than a preset amount of time, the AWS will return to its pre-hostilities setting: permitted to fire when allowed under the *Caroline* standard, but not to engage targets of opportunity. Fourth, AWS should be subject to an "orange box" rule: these systems should be required to store and transmit as much data as possible regarding its decision to strike in anticipatory self-defense.

The development and introduction of AWS may be inevitable, but humans retain decision-making power over one important set of rules: under what circumstances may autonomous systems take us to war? Even if one disagrees with the analysis and proposed rules presented here, that inquiry deserves an answer. We cannot afford AWS that shoot first and ask these questions later.