INTERNATIONAL LAW STUDIES

Published Since 1895

Autonomous Cyber Capabilities Below and Above the Use of Force Threshold: Balancing Proportionality and the Need for Speed

Peter Margulies

96 INT'L L. STUD. 394 (2020)

Volume 96 2020

Published by the Stockton Center for International Law

ISSN 2375-2831

Autonomous Cyber Capabilities Below and Above the Use of Force Threshold: Balancing Proportionality and the Need for Speed

Peter Margulies[*]

CONTENTS

[*] Professor of Law, Roger Williams University School of Law. B.A., Colgate; J.D., Columbia Law School. I thank Gary Brown, Tim McFarland, and Kurt Sanger for comments on a previous draft.

This article originated from a NATO Cooperative Cyber Defence Centre of Excellence project examining autonomous cyber capabilities. It and other papers produced during the project will appear in AUTONOMOUS CYBER CAPABILITIES UNDER INTERNATIONAL LAW (Rain Liivoja & Ann Väljataga eds. 2021).

The thoughts and opinions expressed are those of the author and not necessarily those of the U.S. government, the U.S. Department of the Navy, or the U.S. Naval War College.

I. INTRODUCTION

In the fragile domain of computer network security, seconds can mean the difference between responding effectively to an incursion and sustaining devastating damage. Using those precious seconds is a job for machines, not humans. An autonomous computer system—defined as software that chooses particular actions without specific human pre-approval—can respond quickly.[1] However, reliance on machines has its perils, including ensuring that machines that go beyond mere defense do so in compliance with applicable international law principles such as proportionality.[2] Autonomous systems' flaws—brittleness, bias, and unintelligibility—compound these challenges.

Solving those challenges is important since proportionality is central in several contexts. First, in the *jus ad bellum*, self-defense must be tailored to the goal of stopping an adversary's attacks.[3] Second, in the *jus in bello*, the rule

[1] *See* TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS 128 (Michael N. Schmitt ed., 2d ed. 2017) [hereinafter TALLINN MANUAL 2.0] (noting the speed of cyber exchanges); UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH, THE WEAPONIZATION OF INCREASINGLY AUTONOMOUS TECHNOLOGIES: AUTONOMOUS WEAPON SYSTEMS AND CYBER OPERATIONS 4 (2017), https://www.unidir.org/files/publications/pdfs/autonomous-weapon-systems-and-cyber-operations-en-690.pdf (noting that as part of the 2015 "Grand Cyber Challenge" competition, the U.S. Department of Defense's Defense Advanced Research Projects Agency sought "[m]achines . . . to find and patch [software flaws] within seconds . . . and find their opponents' weaknesses"); *see also* ROBIN GEISS, THE INTERNATIONAL-LAW DIMENSION OF AUTONOMOUS WEAPONS SYSTEMS 9 (2015), http://library.fes.de/pdf-files/id/ipa/11673.pdf (noting that the U.S. National Security Agency is allegedly working on software that will autonomously analyze data inputs and when necessary respond to cyber attacks from abroad); *cf.* PAUL SCHARRE, ARMY OF NONE: AUTONOMOUS WEAPONS AND THE FUTURE OF WAR 214–16 (2018) (describing autonomous features of Stuxnet, a means allegedly designed and deployed by the United States and Israel to insert a software flaw into the industrial control systems running centrifuges that were part of the Iranian nuclear program).

[2] Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 JOURNAL OF NATIONAL SECURITY LAW AND POLICY 1 (2019); *see also* Alan Schuller, *At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapons Systems with International Humanitarian Law*, 8 HARVARD NATIONAL SECURITY JOURNAL 379 (2017) (discussing autonomous systems and law of armed conflict).

[3] *See* TALLINN MANUAL 2.0, *supra* note 1, at 349.

of proportionality means that the harm to civilians expected cannot be excessive in light of the military advantage that the planner anticipates.[4] Third, a State's countermeasure in response to a violation of its sovereignty or a breach of the principle of nonintervention should center on persuading the responsible State to comply with its obligations.[5] This article also argues that the duty to take feasible precautions—expressly stated in the *jus in bello*[6]—is inherent in *all* proportionality requirements, including those governing the *jus ad bellum* and countermeasures.

Proportionality serves vital purposes. In the *jus in bello*, it limits harm to key interests, including the liberty, safety, and welfare of civilians and the integrity of civilian infrastructure. Proportionality in countermeasures also safeguards State interests, curbing a victim State's impingements on a responsible State's sovereignty after an incursion that may have involved limited impact. In requiring some fit between a response and an initial incursion, proportionality in the *jus ad bellum* and in countermeasures limits escalation.

[4] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 51(5)(b), June 8, 1977, 1125 U.N.T.S. 3 [hereinafter Additional Protocol I].

[5] *See* TALLINN MANUAL 2.0, *supra* note 1, at 128; Air Service Agreement of 27 March 1946 (U.S. v. Fr.), 18 R.I.A.A. 417, 443, ¶ 83 (Perm. Ct. Arb. 1978); Michael N. Schmitt, *"Below the Threshold" Cyber Operations: The Countermeasures Response Option and International Law*, 54 VIRGINIA JOURNAL OF INTERNATIONAL LAW 697, 715 (2014). This article takes no position on whether respect for sovereignty per se is part of the backdrop of international law or instead constitutes a primary rule. *Compare* Michael N. Schmitt & Liis Vihul, *Respect for Sovereignty in Cyberspace*, 95 TEXAS LAW REVIEW 1639, 1644–49 (2017) (suggesting that prohibition on violations of sovereignty, particularly through incursions on territory of another State short of the actual use of force constitutes a primary rule of international law), *with* Gary P. Corn & Robert Taylor, *Sovereignty in the Age of Cyberspace*, 111 AMERICAN JOURNAL OF INTERNATIONAL LAW 207, 209–10 (2017) (arguing that respect for sovereignty is an overarching principle, rather than a basis for a separate rule barring incursions on sovereign territory when those incursions are too fleeting or marginal to constitute a use of force); *see also* Jeremy Wright, U.K. Attorney General, Address at Chatham House: Cyber and International Law in the 21st Century (May 23, 2018), https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century (agreeing that respect for sovereignty per se is not a "specific rule" that exceeds the scope of the principle of nonintervention); *cf.* Eric Talbot Jensen, *The* Tallinn Manual 2.0*: Highlights and Insights*, 48 GEORGETOWN JOURNAL OF INTERNATIONAL LAW 735, 741–42 (2017) (taking a middle position in the sovereignty per se debate, especially regarding the content of rules on territorial incursions below the use of force threshold, and arguing that the question turns on the "domain and practical imperatives of states").

[6] Additional Protocol I, *supra* note 4, art. 57(2)(a)(ii); Geoffrey S. Corn, *War, Law, and the Oft Overlooked Value of Process as a Precautionary Measure*, 42 PEPPERDINE LAW REVIEW 419, 459 (2015).

In the *jus ad bellum* and countermeasures contexts, this article argues that a victim State should receive a "margin of appreciation"—a measure of deference—in crafting an answer to incursions by a responsible State.[7] An unduly strict reading of proportionality can stifle victim States' responses, creating a "first-mover" advantage when a State uses force unlawfully[8] or breaches the principle of non-intervention.[9] Regaining the initiative for victim States is particularly pressing in the cyber realm, where an initial attack can occur with great speed while engendering broad effects. Autonomous cyberagents can provide that necessary response capability.[10]

An autonomous cyber agent is software that designers have set up to make and execute decisions without prior approval from human beings. Even though autonomous agents can act with a speed that humans cannot match, serious flaws mar autonomous agents' performance. Autonomous agents lack contextual judgment and their reasoning can be "brittle," since changing details in their inputs can spur arbitrary changes in outputs.[11] Autonomous models' outputs can also be biased due to skewed inputs or faulty

[7] The European Court of Human Rights has granted States a margin of appreciation in tailoring individual rights such as the right of free expression to each State's society and culture. *See* Zana v. Turkey, App. No. 18954/91, ¶ 51(ii) (1997) (ECtHR), http://hudoc.echr.coe.int/eng?i=001-58115. I suggest in this article that a similar concept has informed development of the law of countermeasures, providing a victim State with a measure of flexibility—albeit flexibility within reasonable bounds—in crafting proportional countermeasures. *See Air Service Agreement*, *supra* note 5, ¶ 83 (suggesting that assessing proportionality in countermeasures necessarily involved an "approximation" of the scale of action by the victim State to induce the responsible State to comply with its obligations).

[8] *See* TALLINN MANUAL 2.0, *supra* note 1, at 331–35 (explaining that use of force in the cyber realm entails effects that are akin to kinetic actions in severity, immediacy, directness, and invasiveness).

[9] Some commentators have argued that the law of countermeasures is unduly restrictive, unduly burdening victim States' responses. *See* Gary Corn & Eric Jensen, *The Use of Force and Cyber Countermeasures*, 32 TEMPLE INTERNATIONAL AND COMPARATIVE LAW JOURNAL 127 (2018). Ensuring that a countermeasure is an effective remedy—especially in the cyber domain—may require some streamlining and modest revision of legal requirements. For example, because time may be of the essence, States need flexibility in determining whether to provide a responsible State with notice of a pending countermeasure. *See* TALLINN MANUAL 2.0, *supra* note 1, at 120.

[10] *See* Deeks, Lubell & Murray, *supra* note 2, at 7–8.

[11] Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUMBIA LAW REVIEW 1851, 1877–78 (2019).

inferences from that data.[12] In addition, those outputs can be unexplainable, crunching myriad variables in counterintuitive ways.[13] Moreover, "automation bias" leads humans to exaggerate the accuracy of technology, even as technology often fails to deliver on its promise.[14]

To address the flaws of autonomous cyberagents that would otherwise undermine compliance with the proportionality principle in international law, this article suggests that States must also take all feasible precautions to reduce harm to civilians, civilian objects, and sovereign interests. While international law makes this duty express in the *jus in bello*, this article argues that the duty to take due care through feasible precautions is present in both the *jus ad bellum* and countermeasures.[15] That duty is both substantive and evidentiary, with the substantive duty representing *lex ferenda*, and evidentiary component constituting *lex lata*.

As a substantive matter, both the *jus ad bellum* and the law of countermeasures are gradually moving toward an acknowledgment that due care is a component of proportionality, at least in the interdependent cyber domain. Reflecting this emerging duty, when a State engages in lawful self-defense against another State, unintended spillover effects on a third State's networks arising from that action would constitute an unlawful use of force.[16] Feasible precautions that would reduce that spillover are a logical implication of the *jus ad bellum*'s prohibition on the use of force. Similarly, the "interdependent nature" of networks makes due care a component of proportionality in countermeasures.[17]

In addition, feasible precautions are important from an evidentiary perspective as *proof* that a State has exercised the due care that proportionality requires. If a State has made feasible efforts to reduce the consequences of

[12] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROCEEDINGS OF MACHINE LEARNING RESEARCH 1 (2018), https://perma.cc/8CX2-AMWM.

[13] Zachary C. Lipton, *The Mythos of Model Interpretability*, at 4, ARXIV LABS, https://arxiv.org/pdf/1606.03490.pdf (last revised Mar. 6, 2017).

[14] *See* Claudia E. Haupt, *Artificial Professional Advice*, 21 YALE JOURNAL OF LAW AND TECHNOLOGY 55, 71 (2019) (noting that humans reviewing an agent's work—such as a medical diagnosis based on radiological imaging—may do only a cursory job because they believe the agent is virtually always correct).

[15] Consistent with this implication, *Tallinn Manual 2.0* finds that States considering countermeasures must "exercise considerable care" in ensuring compliance with proportionality. *See* TALLINN MANUAL 2.0, *supra* note 1, at 128.

[16] *Id.* at 333–34.

[17] *Id.* at 128.

its actions in cyberspace, external audiences—be they other States, tribunals, scholars, or nongovernmental organizations—will be more likely to find that any effects beyond strict proportionality are *de minimis*. To codify this natural tendency, feasible cyber precautions in the *jus ad bellum* and countermeasures should be regarded as a prerequisite for the margin of appreciation that a target State enjoys in these arenas.

Because of the need for speed in the cyber realm, the timing of feasible precautions will be more flexible than in kinetic operations. In some cases, a State may need to plan feasible precautions *before* a specific incursion from another State. That is particularly important if the feasible precaution involves training an autonomous agent. For example, gathering intelligence about an adversary may be a necessary component of such training. In addition, while a State may need to respond quickly to another State's cyber incursion, a State that has responded may be able to reduce needless damage *after* its response through assisting in the repair of damaged networks. The timing of required feasible precautions should be sufficiently flexible to include both pre-incursion and post-response measures.

This article groups feasible precautions into four categories: reconnaissance, coordination, repairs, and review. Reconnaissance entails efforts to map an adversary's network in advance of any incursion by that adversary, since time may be elusive after an incursion.[18] On this view, acts of cyber espionage, such as the use of honey pots, are not merely permitted, but *required*, at least if they are feasible.[19] Coordination requires that a cyberagent rely on more than one algorithm, machine, or sensor; often it will entail the interaction of multiple systems, including one or more that will keep watch on the primary agent. In addition, a State must, where feasible, assist in the repair of damage it has caused through a countermeasure, including secondary effects felt by third-party States. Where a responding State can provide a patch to address secondary effects, if feasible, it must. In the *jus in bello*, patching should reduce the net quantum of harm to civilian persons or objects

[18] *See id.* at 128 (discussing mapping as a prelude to countermeasures, while also suggesting that mapping will typically occur after the responsible State's breach of duty that occasioned the possible countermeasure).

[19] A honey pot is a decoy file or other data asset that an entity designs to attract intruders and detect intrusions. If a user who has not received authorization gains access to the honey pot, the entity that set up the decoy can detect that unauthorized use. In addition, the entity can learn other information about the intruder's methods that can also be helpful in conducting future probes of the intruder. *See* Amanda N. Craig, Scott J. Shackelford & Janine S. Hiller, *Proactive Cybersecurity: A Comparative Industry and Regulatory Analysis*, 52 AMERICAN BUSINESS LAW JOURNAL 721, 756 n.147 (2015).

ascribed to the attack in the proportionality calculus, and should play a similar role in the *jus ad bellum* and countermeasures. Finally, planners must regularly review the performance in the field of autonomous cyberagents.

These precautions will not ensure compliance with the principle of proportionality in all cases involving autonomous cyberagents. But they will both promote compliance and provide States that take these precautions with a limited safe harbor, a margin of appreciation for effects that would otherwise violate the duty of proportionality in the *jus ad bellum* and countermeasures. In the *jus in bello*, taking the measures described above would comply with the rule of precautions in attack.

This article proceeds as follows. Part II discusses cybersecurity and then segues into an in-depth account of autonomy, including both virtues, such as speedy response and analysis of multiple variables, and flaws, including unintelligibility, brittleness, and bias. Part III notes the principle of distinction and the rule of proportionality in the *jus in bello*, and then analyzes in greater detail the role of proportionality in the *jus ad bellum* and countermeasures. It also discusses the rule of precautions in its express status under the *jus in bello* and its implied function as a component of proportionality in the other two bodies of law discussed here. Part IV discusses the categories of precautions outlined here: reconnaissance, coordination, repair, and review. This approach will maximize autonomy's tactical strengths in the cyber arena while curbing the effects of autonomy's flaws. Part V concludes.

II. TWO CHALLENGING TECHNOLOGICAL ARENAS: CYBER AND AUTONOMY

Both the cyber domain and autonomous systems feature new technological challenges and capabilities.[20] This Part briefly outlines these issues. It stresses challenges facing autonomous systems to highlight the importance of legal rules to govern autonomous agents.

A. Cyber Incursions: The Turn to a More Proactive Response

The world increasingly relies on computer networks and the Internet for information, communication, and even acquiring essential goods and services. Without the Internet, both daily life and everyday governance would be far

[20] On the challenges posed by new technologies, *see* Eric Talbot Jensen, *The Future of the Law of Armed Conflict: Ostriches, Butterflies, and Nanobots*, 35 MICHIGAN JOURNAL OF INTERNATIONAL LAW 253, 257–58 (2014).

more difficult. As a result of this dependence, incursions on the Internet have taken center stage both in global affairs and military planning.[21]

These incursions have taken a variety of forms. Among the most common are distributed denial of service (DDoS) attacks, where an actor uses masses of computers (botnets) to deluge websites with email or other communications, effectively rendering those sites dysfunctional for some period of time.[22] States and non-State actors also can launch malicious software (malware) that can exfiltrate data for purposes of identity theft, pilfering of intellectual property, or espionage.[23] In another type of incursion, States and others can use malware to manipulate software or destroy data stored on other networks.[24] In incursions such as Stuxnet (sometimes called Olympic Games), States or other actors can manipulate software to compromise industrial control systems (ICS), causing kinetic damage.[25] For example, Russia launched coordinated information operations that used thousands of computers to impersonate persons and groups on social media, spread misinformation, and influence democratic elections, most notably the 2016 U.S. presidential campaign.[26]

[21] *See* U.S. CYBERSPACE SOLARIUM COMMISSION, REPORT 8–16 (2020), https://www.solarium.gov/.

[22] *See* David A. Wallace & Christopher W. Jacobs, *Conflict Classification and Cyber Operations: Gaps, Ambiguities, and Fault Lines*, 40 UNIVERSITY OF PENNSYLVANIA JOURNAL OF INTERNATIONAL LAW 643, 652 (2019).

[23] *See* U.S. CYBERSPACE SOLARIUM COMMISSION, *supra* note 21, at 8–9.

[24] *See* Dan Effrony & Yuval Shany, *A Rule Book on the Shelf?* Tallinn Manual 2.0 *on Cyberoperations and Subsequent State Practice*, 112 AMERICAN JOURNAL OF INTERNATIONAL LAW 583, 620–23 (2018).

[25] In the Stuxnet episode, two States—reportedly the United States and Israel—introduced malware into the ICS that ran the centrifuges used to process uranium for Iran's nuclear program. As a result, the centrifuges overheated and had to be replaced, requiring much time, effort, and expense that set back the Iranian nuclear program. *See* Wallace & Jacobs, *supra* note 22, at 655–56.

[26] *See* U.S. CYBERSPACE SOLARIUM COMMISSION, *supra* note 21, at 68; Michael N. Schmitt, *"Virtual" Disenfranchisement: Cyber Election Meddling in the Grey Zones of International Law*, 19 CHICAGO JOURNAL OF INTERNATIONAL LAW 30 (2018); Sean Watts & Theodore T. Richard, *Baseline Territorial Sovereignty and Cyberspace*, 22 LEWIS & CLARK LAW REVIEW 771, 790 (2018) (discussing Russian efforts to use human trolls to influence Ukrainian elections); *cf.* Effrony & Shany, *supra* note 24, at 609–11 (discussing Russian hacking of the U.S. Democratic Party as part of its election influence operations).

States have sought to develop timely and effective responses to these incursions. For example, the United States recently outlined a "defend forward" component of its "persistent engagement" strategy.[27] That strategy heralds a more proactive approach to parrying cyber incursions. As part of that strategy, U.S. cyber forces temporarily deprived a Russian government unit, the Internet Research Agency, of access to the Internet during the 2018 U.S. election.[28] That visible U.S. response is a powerful signal that victim States will not remain passive in the face of cyber incursions.

B. *Autonomy*

Since the broad outlines of cybersecurity have become widely known, this Section discusses the cyber domain's companion: autonomy. It starts with a brief account of modes of autonomy, defined as artificial intelligence making substantive decisions that affect commerce, industry, domestic governance, and both conflict and competition between States—in other words, a broad swath of "human" endeavor. Further, the following subsection notes problems with autonomy, including brittleness, bias, and unintelligibility.

1. Modes of Autonomy

Autonomy involves models of artificial intelligence that draw inferences, discern patterns, and initiate actions based on machine learning.[29] A computer designer trains the machine agent or "learner" on a large quantity of data, called a "training set." The designer then tests the agent with a "test set" of

[27] *See* U.S. DEPARTMENT OF DEFENSE, CYBER STRATEGY 2018: SUMMARY (2018), https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF; *see also* U.S. CYBERSPACE SOLARIUM COMMISSION, *supra* note 21, at 33–34.

[28] *See* Erica Borghard, *Operationalizing Defend Forward: How the Concept Works to Change Adversary Behavior*, LAWFARE (Mar. 12, 2020), https://www.lawfareblog.com/operationalizing-defend-forward-how-concept-works-change-adversary-behavior.

[29] PEDRO DOMINGOS, THE MASTER ALGORITHM (2015); STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (3d ed. 2010); Peter Margulies, *Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts*, *in* RESEARCH HANDBOOK ON REMOTE WARFARE 405, 415–31 (Jens David Ohlin ed., 2017); Emily Berman, *A Government of Laws and Not of Machines*, 98 BOSTON UNIVERSITY LAW REVIEW 1277, 1286–90 (2018); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC DAVIS LAW REVIEW 653 (2017).

data inputs to determine if the training has enabled the agent to draw inferences or assess patterns with sufficient precision to merit deployment in a particular task. Once the agent has performed the task in the field for some period, the designer reviews the agent's performance to determine if modifications to the agent's training are necessary.

There are many models of machine learning. Two useful examples are decision trees and artificial neural networks. Each is addressed in turn.

Decision trees analyze data points in a choice between several possible actions.[30] The decision tree shows the interplay of those factors in graphic terms, like leaves or branches in a diagram. Each leaf stands for a specific data point that has played a role in a particular decision.[31] In a classic example, a decision by a group of friends on waiting for a dinner table at a crowded restaurant might depend on factors such as the day of the week (on weekends, other restaurants might also be packed), the style of cuisine served, the time of the friends' last meal, the weather (torrential rain would strengthen the case for staying put), and the comfort and availability of the restaurant's bar (which could blunt the pain of waiting). To manage the large amounts of data that disparate variables can produce, a decision tree will generate an explanation that is as simple as possible, given the data, with leaves pruned away if they are unnecessary for prediction. For example, suppose that our prospective diners cared less about the style of food than they did about their ability to amble up to the bar and secure the beverage of their choice. In that event, an autonomous agent would prune away the leaf representing the restaurant's high culinary rating. Because of its leaves, a decision tree's reasoning is intelligible, facilitating review of inputs and outputs.

While artificial neural networks lack the graphic, readily accessible outputs of decision trees, they are often more accurate than decision trees in detecting patterns in masses of data, including information from other autonomous agents. The structure of neural networks resembles the human brain,[32] which consists of neurons that are connected with countless vanish-

[30] This discussion relies on Peter Margulies, *Surveillance by Algorithm: The NSA, Computerized Intelligence Collection, and Human Rights*, 68 FLORIDA LAW REVIEW 1045, 1063–71 (2016); *see also* Shin-Shin Hua, *Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control*, 51 GEORGETOWN JOURNAL OF INTERNATIONAL LAW 117, 124–26 (2019) (discussing models of machine learning); *see generally* RUSSELL & NORVIG, *supra* note 29.

[31] RUSSELL & NORVIG, *supra* note 29, at 757.

[32] *Id.* at 728.

ingly thin fibers. Neurons in an artificial network are connected through layers that perform particular facets of a task. Through that layered structure, the layers break down masses of data into manageable steps.

Recently, research on neural nets has focused on "deep learning," in which designers use many layers to analyze a broad spectrum of variables, sometimes called "dimensions." While humans are limited in the dimensions they can grasp, machines do not labor under such constraints. For humans, scenes in everyday life occur in three dimensions: length, breadth, and depth. Pictorial representations occur in two dimensions—height and width—although they may provide the illusion of depth. Similarly, most graphs plot points at the intersection of two axes, such as time and some measure of quantity or frequency.[33] For human beings who cannot visualize beyond three dimensions, a deep learning agent's inputs and outputs can be difficult to depict visually and grasp in operational terms. A graph of a deep learner's outputs could have ten or twenty axes, instead of the two (one horizontal and one vertical) that appear in most graphs. Depicting that complex, interactive set of outputs in a fashion that humans can understand poses a special challenge.

As an example, consider how a neural net would perform an increasingly common task for artificial intelligence: facial recognition. After designers inputted a very large training set, a neural net would analyze an even larger collection of photographs or videos depicting the human face in a range of angles and contexts. A neural net would subdivide its task, first searching for faces within a variety of objects and shapes in each photograph, which might also include full- or partial-body views, inanimate objects such as vehicles, and other living things such as trees or flowers. Other layers of the neural net would search for facial features, such as eyes and noses. Another layer would flag specific faces.[34]

[33] *See* Geoffrey E. Hinton, *Learning to Represent Visual Input*, 365(1537) PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B: BIOLOGICAL SCIENCES, Jan. 12, 2010, at 177, 180–83.

[34] *See* Lucas D. Introna & Helen Nissenbaum, *Facial Recognition Technology: A Survey of Policy and Implementation Issues* 17–18 (Lancaster University Mgmt. Sch., Working Paper No. 2010/030, 2010), https://eprints.lancs.ac.uk/id/eprint/49012/. The description in the text aims to explain how a neural net's hidden layers work. It does not purport to offer an up-to-date account of current developments in facial recognition technology, which are beyond the scope of this article.

At their best, autonomous agents can draw inferences quickly from masses of data that would take humans months or years to digest.[35] Moreover, autonomous agents can spot patterns that would escape human detection. But autonomous agents have several significant flaws that require human acknowledgment and attention. These flaws are the subject of the next subsection.

2. Autonomy's Liabilities

Despite their prodigious achievements and extraordinary promise, autonomous agents also suffer from significant deficits. For example, autonomous agents' reasoning is "brittle" and lacks contextual judgment, prompting major mistakes based on minor changes in inputs. Autonomous models can also exhibit bias, either because of problems with the inputs they receive from human developers in the course of training or because of flaws in their analysis of those inputs. In addition, autonomous agents produce outputs that are sometimes unintelligible or unexplainable, at least in conventional verbal terms. Moreover, an additional bias complicates all human efforts to review the outputs of autonomous agents: humans suffer from "automation bias"—a tendency to exaggerate the accuracy of technology.[36] I discuss each in turn.

a. Autonomy's Brittle Disposition

Many machine learning systems, such as neural networks, are brittle. While their outputs are often highly accurate, they can also make serious mistakes because of seemingly minor changes to the inputs they receive.[37] Autonomous agents make such mistakes because they lack the contextual backdrop

[35] *See* Ian S. Henderson, Patrick Keane & Josh Liddy, *Remote and Autonomous Warfare Systems: Precautions in Attack and Individual Accountability*, *in* RESEARCH HANDBOOK ON REMOTE WARFARE, *supra* note 29, at 335, 341.

[36] *See* Haupt, *supra* note 14, at 71.

[37] Bita Dervish Rouhani et al., *Safe Machine Learning and Defeating Adversarial Attacks*, 17 IEEE SECURITY AND PRIVACY, Mar.-Apr. 2019, at 31, 31–32, https://ieeexplore.ieee.org/document/8677311 (noting that neural networks can mischaracterize minor changes in inputs, such as specks at the edges of a photo, as changing the actual subject of the photo, for example, treating a stop as a yield sign); Douglas Heaven, *Why Deep-Learning AIs Are so Easy to Fool*, NATURE, Oct. 9, 2019, https://www.nature.com/articles/d41586-019-03013-5 (noting that developers conducting experiments have found that rotating an object in an image so that the image contains a sideways or upside-down view materially reduces an

that even young children possess. Humans understand the context of images or other data, while machines often lack this holistic understanding. For humans, context determines which characteristics are most important in a given situation. For example, most human beings recognize the classic octagonal shape and red color of a stop sign, even if they cannot read the word "Stop" on the sign itself. The sign's distinctive shape and color are part of a human being's background understanding of context, which that person has learned from experience. In contrast, a neural network only knows context through the data that designers have fed it. That incomplete grasp of context can lead to arbitrary and seemingly random results.

The familiar image of a stop sign illustrates the machine's tendency to respond to minor changes in inputs with major, seemingly arbitrary changes in outputs. Consider the problem posed by so-called adversarial examples. Because a neural network's understanding of context has so many gaps, computer scientists have been able to prompt neural networks—including those trained to recognize text—to mistakenly classify a stop sign as a yield sign, simply by superimposing a few small white specks on a stop sign photo.[38] Humans alert to context would not take the bait; they would ignore the specks and focus on salient factors such as shape, color, and text that distinguish the stop sign from other road signs. As another example, a child who sees a photograph of a lion that is upside down will still recognize the image as portraying a lion, because based on his or her experience, the child recognizes salient traits of the lion, such as its mane, teeth, and tail. However, a machine lacks that contextualized library of salient characteristics. For an autonomous agent, rotating an image may make the image appear to be different in kind, rather than merely in format.

An autonomous agent lacks an inherent understanding of the difference between format and content. Children grasp that distinction in grade school when they submit a written homework assignment in a particular font and line spacing. Computers can learn the distinction, but only through exposure to training data. Their sense of context is thin and therefore subject to substantial volatility. That brittleness is a serious flaw.

Developers often underestimate or fail to identify an agent's gaps in contextual understanding. Socrates taught us more than two thousand years ago

agent's ability to identify the object, even when the agent has successfully identified the object in the past).

[38] *See* Rouhani et al., *supra* note 37, at 32; *see generally* SCHARRE*, supra* note 1.

that much of wisdom is knowing what we do not know.[39] But some failures of knowledge are difficult to detect precisely because humans tend to think that their knowledge is comprehensive and has no gaps—they do not know what they do not know. Developers who have spent countless hours training an autonomous agent also tend to overlook gaps in an agent's knowledge. Similarly, humans are not very good at ascertaining what they *know*. Aspects of human understanding—including the mundane distinction between a stop sign and a yield sign—reflect inferences from data inputs that a designer will have to replicate in a machine to ensure that the machine can draw a similar inference.

As an example of this gap between a developer's inflated view of an agent's knowledge and the agent's *actual* knowledge of context, consider the following example from the healthcare field, where the use of autonomous agents has skyrocketed. Designers trained an agent to rank the urgency of treatment for patients with pneumonia, performing medical triage by prioritizing those at higher risk—patients who were likely to become seriously ill more quickly in the absence of treatment. The trained model classified pneumonia patients who also had asthma as being *low* risk, even though common sense would indicate that the combination of asthma—a serious respiratory disease in its own right—and pneumonia is a *high*-risk condition that requires immediate treatment.[40]

In our pneumonia example, factual inputs led the model to incorrectly classify asthmatic pneumonia patients as low risk because doctors know the high risk of this combined condition and hence admit such patients directly to hospital intensive care units where patients promptly receive treatment. That prompt treatment then yields better clinical outcomes. In essence, the model mistook effects for causes, thereby drawing exactly the wrong lesson from the data. Exercising common sense, a layperson would have inferred that favorable outcomes for pneumonia patients with asthma were merely a beneficial *effect* of the accurate judgment that those patients posed a high risk. In contrast, the autonomous agent badly misread this effect of the high risk

[39] *See* James Grimmelmann, *Listeners' Choices*, 90 UNIVERSITY OF COLORADO LAW REVIEW 365, 377 (2019) (discussing the views of ancient Greek philosopher Socrates and modern economist Kenneth Arrow on the consequences of human beings' tendency to exhibit "partial ignorance," including a person's inability to accurately judge what he or she does not yet know).

[40] Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 UNIVERSITY OF CHICAGO LAW REVIEW 1311, 1336–37 (2019); Strandburg, *supra* note 11, at 1877–78.

of asthma as suggesting a causal relationship between the asthma itself and positive clinical outcomes.

Reliance on this model without further investigation would have upended appropriate treatment priorities. Yet researchers discovered this error only by sifting through the model's outputs and comparing them with inputs. Once researchers discovered the error, they could feed the model additional data to properly discount asthma patients' favorable outcomes. However, predicting all these gaps in advance is difficult, if not impossible.

b. Bias and the Challenge of Inputting Optimal Data

Bias is also an issue with machine learning and autonomous agents. Autonomous agents will not exhibit the emotions, such as anger or fear, which contribute to bias in human beings. Nevertheless, bias often seeps into autonomous agents' outputs. That invidious influence has several contributing causes, including, (1) human biases that affect the labeling and type of data that designers use to train agents, and, (2) shortcomings in inferences based on that data that can afflict both humans and machines.[41]

As an example of the first cause of bias, consider supervised learning. Recall that in supervised learning, the machine learns from data that humans have already labeled. Since human beings are biased and often fail to recognize their own biases, the labels created by human data workers may reflect bias along several axes, including race, class, religion, and nationality. Labeled data has attracted negative attention recently because some widely used labeled data sets, such as those for faces, seemingly have reflected the biases of the labelers.[42] Identifying these biased labeled data sets and eliminating their discriminatory effects are key goals of current neural network developers. However, the problem is not that easy to solve. Consider unsupervised

[41] *See* Ashley S. Deeks, *Predicting Enemies*, 104 VIRGINIA LAW REVIEW 1529, 1563–65 (2018); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE LAW JOURNAL 1043, 1080–81 (2019).

[42] *See* Cade Metz, *'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You?*, NEW YORK TIMES, Sept. 20, 2019, https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html; Tom C.W. Lin, *Artificial Intelligence, Finance, and the Law*, 88 FORDHAM LAW REVIEW 531 (2019); *see generally* Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA LAW REVIEW 54 (2019); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GEORGIA LAW REVIEW 109, 133 (2017) (discussing problems of bias and incorrect or incomplete information that skew labeling data sets); Sandra Mayson, *Bias In, Bias Out*, 128 YALE LAW JOURNAL 2218, 2227–38 (2019).

learning, in which designers merely feed the machine vast amounts of unlabeled data, leaving the machine to discern patterns and anomalies on its own. Here too, the types of data fed to the machine, as well as the ratio of different inputs, can reflect human biases.[43]

The brittleness noted above can also foment bias. Suppose that agents trained to detect or monitor terrorism settle on superficial commonalities of some terrorists. For example, male terrorists who profess to follow Islam may well have facial hair, since that is one aspect of Islamic religious observance. That attribute obviously is an inaccurate metric for predicting terrorism. It is underinclusive: domestic terrorism in the United States stems from white nationalist groups or others who are not Muslim and may be less likely to sport facial hair.[44] The facial hair metric is also overinclusive: terrorism is a low-incidence event, and the overwhelming majority of male Muslims with facial hair do not commit acts of terrorism.[45] Reliance on the presence of facial hair is thus a poor metric for identifying terrorists. But a machine may not sort through all the data necessary to draw that conclusion, or may not receive that data from its human designers.

The tendency to express bias can easily infiltrate the cyber realm. Autonomous cyberagents that have access to analyses of likely cyber culprits may infer that cyber threats come from particular States, such as Iran, which also trigger concern about terrorist threats. In a particular case, that inference may lead to the correct result. But a focus on Iran as the cyber culprit of choice would be markedly underinclusive. Cyber incursions also emanate from a host of other sources, including States such as China and Russia, non-State armed groups, and criminal organizations.[46] The information available to the agent may be incomplete or biased. It may then lead to mistaken attributions by the agent, which in turn help drive errant autonomous responses to actual or potential cyber incursions.

[43] *See* Mayson, *supra* note 42, at 2260.

[44] *Cf.* Shirin Sinnar, *The Lost Story of Iqbal*, 105 GEORGETOWN LAW JOURNAL 379 (2017) (discussing effect of stereotypes on the pattern of post-9/11 immigration detention).

[45] *Cf.* Emily Berman, *The Paradox of Counterterrorism Sunset Provisions*, 81 FORDHAM LAW REVIEW 1777, 1801 (2013) (summarizing social science evidence showing that terrorist acts are rare, "low-probability" events).

[46] *See* U.S. CYBERSPACE SOLARIUM COMMISSION, *supra* note 21, at 10–14.

c. Explainability

The most accurate autonomous agents also are the most difficult to explain. Neural networks are generally more accurate than other agents, such as decision trees. But the engine of their accuracy also hinders their explainability. Neural networks are accurate because they use layers to sift through a wide array of variables. Reducing the outputs of that layered analysis into a conventional verbal explanation is difficult.[47] That difficulty can conceal the causes of an inaccurate or biased output.[48] There are several methods for understanding that causation, including posing counterfactuals to the machine and seeing if they change the outputs.[49] Still, explainability poses a vexing issue for autonomous agents, and while scientists are working to develop approaches to reaching this goal, challenges remain.

d. Automation Bias: The Machines Know Best

A review of an agent's outputs is also difficult because of humans' automation bias.[50] In both operational interfaces with agents in the course of a mission and review of the mission's effectiveness, humans tend to defer unduly to machines.[51] Operating in exigent settings such as piloting aircraft, human collaborators with autonomous agents tend to become complacent about the

[47] *See* Lipton, *supra* note 13, at 4; Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 HASTINGS LAW JOURNAL 1389, 1439 (2019) (noting that the number of variables that neural networks process impedes verbal explanations); *see also* DAVID FREEMAN ENGSTROM ET AL., *GOVERNMENT BY ALGORITHM:* ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 28–29 (2020), https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf (suggesting that government enforcement personnel using machine learning to spot illegal conduct in complex arenas such as securities markets may ask what inputs have prompted an autonomous agent to flag a particular individual or firm as a risk); *cf.* Lehr & Ohm, *supra* note 29, at 692 (noting that a developer can often discern "how important different input variables were to the predictions generated [made by the agent] and how changes in the input variables tend to be translated into changes in the outcome variable").

[48] DOMINGOS, *supra* note 29, at 239.

[49] In other words, a designer would vary the inputs to the agent, and determine the effect of each such change on the agent's results. *See* Lipton, *supra* note 13, at 6; Lehr & Ohm, *supra* note 29, at 692.

[50] *See* Haupt, *supra* note 14, at 71.

[51] *See* M. L. Cummings, *Human Supervisory Control Challenges in Network Centric Operations*, at 4, HUMAN AND AUTONOMY LABS: PUBLICATIONS (2005), https://hal.pratt.duke.edu/sites/hal.pratt.duke.edu/files/u13/Human%20Supervisory%20Control%20Challenges%20in%20Network%20Centric%20Operations%20.pdf.

agent's performance. If an agent makes a mistake, this complacency hinders the human collaborator's ability to recognize and rectify the error.[52] For example, in the Boeing 737 Max crashes, it appears that the pilots had difficulty in pivoting from reliance on the agent's navigation decisions to regaining human control over the aircraft.[53]

Moreover, evidence suggests that designers of the aircraft's human-machine interface paid insufficient attention to human habits and response times.[54] Similar problems occur with reviews of agents' outputs. Humans may assume that machines have made correct decisions, and may be less willing to probe as deeply as a comprehensive review would require. Relatedly, humans may not design or deploy autonomous agents to provide optimal aid in this task.

C. Summary

Cyber and autonomy are extraordinarily powerful technologies that can improve human performance. But each has vulnerabilities and deficits. Designers of autonomous cyberagents need to be aware of those flaws. In addition, legal regimes for governing autonomous cyberagents have to display similar awareness. The next Part aids in that task by providing an overview of international law principles applicable to the autonomous cyber domain.

[52] *See* Jin Zhou et al., *The Impact of Different Levels of Autonomy and Training on Operators' Drone Control Strategies*, 8(4) ACM TRANSACTIONS ON HUMAN-ROBOT INTERACTION, Oct. 2019, at 22:1, 22:13, https://dl.acm.org/doi/pdf/10.1145/3344276 (discussing human-machine interaction in drone piloting, including how differences in human training appeared to influence human inclination and ability to make real-time adjustments to navigation path determined by automated software).

[53] *See* Chris Hamby, *How Boeing's Responsibility in a Deadly Crash 'Got Buried,'* NEW YORK TIMES, Jan. 20, 2020, https://www.nytimes.com/2020/01/20/business/boeing-737-accidents.html.

[54] *Id.*

III. AUTONOMY, CYBER, AND PRINCIPLES OF INTERNATIONAL LAW

In the realms of cyber and autonomy, international law applies.[55] International law includes proportionality in the *jus ad bellum*, *jus in bello*, countermeasures, and human rights.[56] Cyber and autonomy may require modest revisions in international law rules relevant to kinetic or other means of action and response. Before addressing the need for further elaboration or revision, we first should outline the relevant international law rules. This Part reviews the relevant rules on proportionality in the *jus ad bellum*, countermeasures, and the *jus in bello*. I also address the *jus in bello* rule of precautions in attack and suggest that some version of that rule applies in the *jus ad bellum* and countermeasures.

A. Distinction, Lethal Weapons, and the Cyber Domain

While the analysis of proportionality here does not directly address the core *jus in bello* principle of distinction, clarification of that fundamental principle is a useful first step. The principle of distinction bars the targeting of civilians in an armed conflict.[57] Much of the controversy about autonomy in armed conflict has stemmed from concern that autonomy poses tensions with this principle.[58] Using computers to make targeting decisions with little or no

[55] *See* TALLINN MANUAL 2.0, *supra* note 1, at 127 (discussing the application of the international law of countermeasures to the cyber domain); Harold Hongju Koh, Legal Advisor, U.S. Department of State, Remarks at the USCYBERCOM Inter-Agency Legal Conference: International Law in Cyberspace (Sept. 18, 2012), https://2009-2017.state.gov/s/l/releases/remarks/197924.htm; Brian Egan, *International Law and Stability in Cyberspace*, 35 BERKELEY JOURNAL OF INTERNATIONAL LAW 169, 177 (2017); Paul C. Ney, Jr., General Counsel, U.S. Department of Defense, DOD General Counsel Remarks at U.S. Cyber Command Legal Conference (Mar. 2, 2020), https://www.defense.gov/Newsroom/Speeches/Speech/Article/2099378/dod-general-counsel-remarks-at-us-cyber-command-legal-conference/; Kristen E. Eichensehr, *The Cyber-Law of Nations*, 103 GEORGETOWN LAW JOURNAL 317 (2015); Michael N. Schmitt, *Wired Warfare 3.0: Protecting the Civilian Population During Cyber Operations*, 101 INTERNATIONAL REVIEW OF THE RED CROSS 333, 334 (2019) (noting "broad consensus that IHL . . . applies to cyber operations during an armed conflict").

[56] This article leaves the important issue of proportionality and human rights for another day. *See* Margulies, *supra* note 30.

[57] Additional Protocol I, *supra* note 4, arts. 48, 51(2).

[58] *See* Kenneth Anderson, Daniel Reisner & Matthew Waxman, *Adapting the Law of Armed Conflict to Autonomous Weapons Systems*, 90 INTERNATIONAL LAW STUDIES 386, 401–05 (2014); Marco Sassòli, *Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified*, 90 INTERNATIONAL LAW STUDIES 308

real-time human ability to veto those decisions could result in substantial noncompliance.[59]

For example, suppose that an autonomous agent mistakenly "learned" through inputted data that it was permissible to attack civilians, or drew unreasonable inferences in identifying a civilian as a direct participant in hostilities subject to targeting.[60] Acting on these mistakes would pave the way for

(2014); Michael N. Schmitt & Jeffrey S. Thurnher, *"Out of the Loop": Autonomous Weapons Systems and the Law of Armed Conflict*, 4 HARVARD NATIONAL SECURITY JOURNAL 231 (2013). Critics of the use of autonomous weapons in armed conflict have outlined comprehensive concerns about compliance with the *jus in bello* and have urged a ban on development of such weapons. *See* Christof Heyns (Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions), *Annual Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, ¶ 55, U.N. Doc. A/HRC/23/47 (Apr. 9, 2013) (warning that autonomous agents do not exhibit "compassion"); Peter Asaro, *On Banning Autonomous Weapons Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, 94 INTERNATIONAL REVIEW OF THE RED CROSS 687 (2012) (asserting that use of autonomous agents in targeting during armed conflict may diminish regard for human life). Other scholars have argued that the critics' concerns are misplaced or exaggerated. *See e.g.*, Chris Jenks, *False Rubicons, Moral Panic, and Conceptual Cul-De-Sacs: Critiquing and Reframing the Call to Ban Lethal Autonomous Weapons*, 44 PEPPERDINE LAW REVIEW 1 (2016).

[59] Compliance with the principle of distinction is a more or less pressing issue depending on the precise nature and purpose of the particular system at issue. *See* U.S. Department of Defense, Directive 3000.09, Autonomy in Weapon Systems 13, 14 (2012, incorporating Change 1, May 8, 2017), https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf (defining an autonomous system as one that "once activated, can select and engage targets without further intervention by a human operator" and noting that some systems "allow human operators to override [autonomous] operation"); *see also* Tim McFarland, The Concept of Autonomy 24 (2020) (unpublished paper on file with NATO Cooperative Cyber Defence Centre of Excellence) (discussing conceptions of autonomy). Autonomous weapons do not necessarily target humans and may be stationary and purely defensive in character. For example, the U.S. Navy has long used fixed autonomous weapons to identify and repel enemy missiles approaching naval vessels. *See* INTERNATIONAL COMMITTEE OF THE RED CROSS, AUTONOMOUS WEAPONS SYSTEMS: TECHNICAL, MILITARY, LEGAL AND HUMANITARIAN ASPECTS 65–66 (2014), https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014. These fixed defensive applications do not raise the same concerns as mobile offensive systems about compliance with IHL. The U.S. Navy is also developing autonomous swarming technology for offensive naval operations that the Navy could at some point use for targeting, although the rules for these systems currently require human supervision.

[60] Michael N. Schmitt, *Deconstructing Direct Participation in Hostilities: The Constitutive Elements*, 42 NEW YORK UNIVERSITY JOURNAL OF INTERNATIONAL LAW AND POLITICS 697, 699 (2010); Kenneth Watkin, *Opportunity Lost: Organized Armed Groups and the ICRC "Direct Participation in Hostilities" Interpretive Guidance*, NEW YORK UNIVERSITY JOURNAL OF INTERNATIONAL LAW AND POLITICS 641, 643–44 (2010).

violations of international humanitarian law (IHL). An autonomous agent's unreasonable decision to use lethal force in an armed conflict would constitute a major challenge to IHL's traditional balance of humanity and military necessity.[61]

Analyzing autonomous agents' compliance with IHL in the cyber realm mutes, but does not eliminate the concerns raised by the prospect of agents' violation of the principle of distinction in kinetic operations.[62] Operations in the cyber domain do not entail direct targeting of persons. Accordingly, concerns about the mistaken or unreasonable use of lethal force are less compelling. Still, such concerns are still relevant. Cyber attacks on civilian sites, such as hospitals, schools, or traffic systems, could cause considerable bodily harm to civilians, as well as damage to civilian objects.[63] Any comprehensive legal regime for autonomous cyberagents must address those issues.

B. *Cyber and Proportionality*

Violations of the rule of proportionality by autonomous cyberagents in the *jus ad bellum*, *jus in bello*, and countermeasures contexts can cause harm to civilians or civilian objects that is excessive or simply needless in light of those cyberagents' legitimate purposes. For example, as suggested above, in an armed conflict an autonomous cyberagent may engage in lawful targeting of a software operating system developed for use by an adversary's military, but in the process may also cause damage to different civilian systems that is foreseeable and excessive in light of the military advantage expected from the underlying attack.[64] Similarly, outside armed conflict, an autonomous cyberagent might take a countermeasure in response to another State's interference. However, the countermeasure might entail effects on the adversary State's sovereign rights that were too far reaching to comply with propor-

[61] *See* Michael N. Schmitt, *Military Necessity and Humanity in International Humanitarian Law: Preserving the Delicate Balance*, 50 VIRGINIA JOURNAL OF INTERNATIONAL LAW 795, 796 (2010).

[62] Duncan B. Hollis, *Autonomous Legal Reasoning in International Humanitarian Law*, 30 TEMPLE INTERNATIONAL AND COMPARATIVE LAW JOURNAL 1, 10–11 (2016).

[63] *See* Oona A. Hathaway, Rebecca Crootof, Philip Levitz, Haley Nix, Aileen Nowlan, William Perdue & Julia Spiegel, *The Law of Cyber-Attack*, 100 CALIFORNIA LAW REVIEW 817, 848 (2012).

[64] *See* TALLINN MANUAL 2.0, *supra* note 1, at 128 (noting potential for disproportionate harm in countermeasures caused by the "interconnected and interdependent nature of cyber systems").

tionality. For these reasons, proportionality's impact on the use of autonomous cyberagents matters, even if the core *jus in bello* principle of distinction is not directly in play.

1. *Jus Ad Bellum* Proportionality

Proportionality in the *jus ad bellum* governs a State's use of force in self-defense against an armed attack.[65] In the cyber realm, the State that has suffered an armed attack must first apply the threshold criterion of necessity, asking whether force—as opposed to use of passive means such as firewalls or active measures such as DDoS incursions that do not rise to the level of force—is reasonably *required* to defeat the attack.[66]

Once a victim State has found that the use of force in self-defense is necessary, it assesses proportionality. Proportionality under the *jus ad bellum* has both functional and quantitative aspects. On a functional level, proportionality asks whether a reasonable person would view the "scale, scope, duration, and intensity" of the force used in self-defense as tailored to the prevention of further attacks.[67] Quantitatively, there will often be some relation in scale, duration, and intensity between an armed attack and force used in self-defense.[68] Those planning the use of force should consider both effects on the initial attacker and collateral impacts on other States, entities, and interests.[69]

Assessing collateral impact is crucial for autonomous cyberagents, given the interconnectedness of the Internet.[70] Responses that are necessary and proportionate for a country that has engaged in an armed attack may well be unnecessary and disproportionate if those responses spill over into other

[65] *See id.* at 340–44.

[66] *Id.* at 348–49.

[67] *See id.* at 349 (asserting that proportionality limits responses to those "required to end the situation that has given rise to the right to act in self-defence").

[68] *See* Enzo Cannizaro, *Contextualizing Proportionality:* Jus ad Bellum *and* Jus in Bello *in the Lebanese War*, 88 INTERNATIONAL REVIEW OF THE RED CROSS 779, 784 (2006) (noting that "[a] state acting in self-defence . . . [should] maintain a certain level of correspondence between the defensive conduct and the attack which prompted it").

[69] Military and Paramilitary Activities in and against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. Rep. 14, 93, ¶ 194 (June 27); *id.* at 269–70, ¶¶ 7, 9; 362–70, ¶¶ 201–14 (dissenting opinion by Schwebel, J.).

[70] *See* RAIN LIIVOJA, MAARJA NAAGEL & ANN VALJATAGA, NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE, AUTONOMOUS CYBER CAPABILITIES UNDER INTERNATIONAL LAW (2019), https://ccdcoe.org/uploads/2019/07/Autonomy-in-Cyber-Capabilities-under-International-Law_260619-002.pdf.

countries not responsible for the attack. In the kinetic domain, it often may be relatively straightforward to restrict a kinetic response to a particular country. For example, a missile strike in self-defense by Arcadia could target military objectives in Ruritania, if the latter country had engaged in an armed attack on Arcadia. Arcadia's strike on Ruritania would generally not affect the third country of Pacifica. However, in the interconnected world of the Internet, in which Pacifica individuals and entities may use servers located in Ruritania, such precision can be more difficult to achieve. The result may be serious impacts on the sovereign interests of Pacifica and other third-party States.

On the other hand, the need for speed in the cyber domain may require greater flexibility in defining both necessity and proportionality. In particular, those concepts should not rigidly require the passage of time between an initial attack by the responsible State and the victim State's response. In the cyber realm, a waiting period of hours or even minutes could mean the difference between preserving the victim State's critical infrastructure and leaving the victim State helpless. Suppose that Ruritania launches an all-out cyber attack on Arcadia's power grid. In this situation, Arcadia may lack the time for a digital forensic investigation to determine whether its passive measures, such as firewalls, have thoroughly blocked Ruritania's attack. Similarly, Arcadia may not have time to ponder whether measures below the use of force threshold will persuade Ruritania to cease its attacks.

In other situations, Arcadia may have the necessary time to assess whether passive measures or countermeasures will adequately address the threat. For example, suppose that Arcadian officials detect phishing emails sent by Ruritanian agents to employees who inspect and maintain ICS at an Arcadian power plant. Those phishing emails contain malware that Arcadia believes could disable the plant's ICS and therefore do serious physical damage to the plant's machinery. In this situation, Arcadia will have time to require the power company to conduct a sweep of its network and send out an emergency notice to its employees to apprise them of the threat and refrain from opening messages that seem suspicious. In this situation, the immediate use of force by Arcadia against Ruritania would be neither necessary nor proportionate.

Read against this shifting factual backdrop, *Tallinn Manual 2.0*'s discussion of necessity and proportionality provides victim States with the flexibility they need without giving them unbounded license in their response. While it is true that the *Manual*'s *ad bellum* discussion of the need to assess the efficacy of passive defenses and countermeasures[71] may imply a specific time sequence in which assessment follows attack, that is not the only possible reading of this passage. As we will see in the next Section, in some situations a State should be able to calibrate its autonomous cyberagents to detect an all-out attack and respond accordingly. In these situations, the victim State should be able to flip the conventional time sequence of attack followed by a necessity and proportionality assessment, and instead rely on a prior "beta test" of its passive defenses and active below-the-force-threshold options. In this situation, requiring a victim State to comport with the conventional time sequence might mean that the victim State would lose the ability to respond *at all*—a result that no State would agree to and that international law does not require. *Tallinn Manual 2.0*'s discussion should not be read to mandate this outcome.[72]

In addition, as discussed later in this Section, a victim State is entitled to a measure of deference or a margin of appreciation in responding to a series of "pinprick" attacks in the cyber realm.[73] Consider a series of phishing attacks by Ruritania on various sensitive government agencies in Arcadia. Assume that those attacks could have kinetic consequences if the malware that Ruritania had implanted in its phishing emails had invaded Arcadian government networks. In response, Arcadia would not be limited to individual attacks that mimicked the Ruritanian incursions. Instead, Arcadia would be allowed to use force equal to a discrete increment *beyond* the aggregation of Ruritania's attacks, as long as that additional increment was reasonable. Assuming digital data or intelligence showed that the Ruritanian attacks were related, permitting Arcadia to aggregate the impacts of Ruritania's attacks would be consistent with proportionality.[74]

Permitting an additional increment beyond the cumulative impact of Ruritanian attacks would allow Arcadia to mount a robust response and ensure that Ruritania accrued no lasting tactical or strategic advantage. In contrast, given the interdependent nature of cyber networks, confining Arcadia to a

[71] TALLINN MANUAL 2.0, *supra* note 1, at 349.

[72] If *Tallinn Manual 2.0* were to be read in this narrow way, its guidance would unduly restrict the options available to victim States under the *jus ad bellum*.

[73] TALLINN MANUAL 2.0, *supra* note 1, at 342, 342 n.823.

[74] *Id.* at 342. *But see* YORAM DINSTEIN, WAR, AGGRESSION, AND SELF-DEFENCE 230–31 (4th ed. 2005) (noting that at least one State has taken this aggregate approach, while other authorities believe that pinprick attacks must be escalating in scale to allow a State to go beyond the force necessary to repel any particular attack).

response to individual pinprick attacks or even to a rigidly demarcated aggregate would, in practice, force Arcadia to stay *below* the level of aggregate impacts. Restricting Arcadia to an aggregate would have that practical effect because an attempt to achieve a precise aggregate in an interconnected online world could well overshoot the mark. Allowing a victim State a margin of appreciation beyond the attacking State's aggregate impacts would ensure that the responsible State's violations of international law did not place the victim State at a permanent disadvantage.

But even with the ability to aggregate impacts and a margin of appreciation in that calculation, proportionality would still impose limits on the victim State's cyber response. For example, suppose Ruritania has attacked an ICS in an Arcadian defense plant and damaged plant machinery. Since Ruritania's attack had kinetic consequences, Arcadia could respond in self-defense.

Arcadia's response could include attacks on the ICS of a Ruritania defense plant. To the extent that a quantitative test for *jus ad bellum* proportionality applies, this response would match the Ruritanian incursion. Indeed, an attack on the ICS of *multiple* Ruritanian defense plants would be within Arcadia's margin of appreciation. So would a targeted temporary power outage or a cyber takedown limited to the Ruritanian military.

However, without a broader Ruritanian attack, an Arcadian response that aimed to destroy the Ruritania power grid as a whole would be disproportionate. Such a response would exceed any quantitative test for *just ad bellum* proportionality and also go beyond what was reasonably necessary to deter further attacks. This reading of the *jus ad bellum* proportionality principle would limit escalation and keep disputes, to the extent possible, within the cyber realm, thereby curbing spillover into the kinetic realm.

2. Proportionality and Countermeasures

This brings us to proportionality in countermeasures. Countermeasures are responses by a victim State to another State's violations of international law.[75] Typically, countermeasures are temporary[76]—a factor that this article views as related to proportionality. Moreover, countermeasures have often entailed notice to the responsible State, although the notice requirement is flexible

[75] TALLINN MANUAL 2.0, *supra* note 1, at 116–17.

[76] *Id.* at 119.

enough to respond to the dictates of practicality.[77] Under current understandings of international law, countermeasures are not available against a non-State actor. Still, a State can target civilian networks—subject to proportionality requirements—in the interest of persuading the responsible State to desist.[78] Countermeasures are not available in collective self-defense, and must be below the level of an armed attack.[79]

In international law regarding countermeasures, proportionality takes into account both a functional aspect—the role of the countermeasure in inducing the responsible State to "comply with its obligations"—and a quantitative aspect—matching the countermeasure with the importance, scale, and duration of the initial action that prompted the countermeasure.[80] More than in the *jus ad bellum*, function and fit are independent criteria. That is, a given countermeasure may be unlawful because it exceeds the importance of the initial action—including its impact on sovereignty—as well as the initial action's scale and duration, *even though* the countermeasure was necessary to induce the responsible State to fulfill its duties.[81]

At the same time, a key arbitral decision on countermeasures recognizes that the fit of a countermeasure need not be precise down to the last decimal point.[82] As the arbitral tribunal noted in the *Air Service* case, "judging the 'proportionality' of counter-measures is not an easy task and can at best be accomplished by approximation."[83] In practice, the willingness to engage in

[77] *Id.* at 120.

[78] *Id.* at 112–13.

[79] *Id.* at 125–26. Many experts believe that a State cannot employ countermeasures above the threshold for the use of force. *Id.* Most States place the use of force at a lower threshold than an armed attack, although the United States believes the two are identical. *Id.* at 126. Countermeasures also may not violate fundamental human rights or *jus cogens*. *Id.* at 123; Rebecca Crootof, *International Cybertorts: Expanding State Accountability in Cyberspace*, 103 CORNELL LAW REVIEW 565, 577–78 (2018).

[80] *See* TALLINN MANUAL 2.0, *supra* note 1, at 128; *Report of the International Law Commission to the General Assembly*, 56 U.N. GAOR Supp. No. 10, at 135, cmt. ¶ 6, U.N. Doc. A/56/10 (2001), *reprinted in* [2001] 2 Yearbook of the International Law Commission, cmt. at 135, ¶ 6, U.N. Doc. A/CN.4/SER.A/2001/Add.1 (Part 2) [hereinafter *Draft Articles of State Responsibility*]; *Air Service Agreement*, *supra* note 5, at 443–44, ¶ 83; Schmitt, *supra* note 5, at 715.

[81] *See Draft Articles of State Responsibility*, *supra* note 80, at 135, cmt. ¶ 7 (noting that "in every case a countermeasure must be commensurate with the injury suffered, including the importance of the issue of principle involved . . . partly independent of the question whether the countermeasure was necessary to achieve the result of ensuring compliance").

[82] *Air Service Agreement*, *supra* note 5, at 443–44, ¶ 83; *Draft Articles of State Responsibility*, *supra* note 80, at 134, cmt. ¶ 3.

[83] *Air Service Agreement*, *supra* note 5, at 443–44, ¶ 83.

approximation means that the victim State receives a measure of deference—in international law, what is often called a margin of appreciation as referred to earlier[84]—in crafting a countermeasure.[85]

The *Air Service* arbitral decision illustrates how this margin of appreciation works. In *Air Service*, against the backdrop of an international air transport compact between France and the United States, France refused to let a U.S. air carrier downsize to a smaller plane in London for a flight that originated in San Francisco and ended in Paris. France alleged that downsizing in the territory of a third State was contrary to the agreement. In response to the French refusal, the United States threatened to block *all* Air France flights from Paris to Los Angeles. The French then agreed to let the flights to Paris continue pending the arbitral decision. The arbitral tribunal ruled that the U.S. threat to discontinue Air France flights from Paris to Los Angeles was not "clearly disproportionate," even though the U.S. threat involved *all* air traffic, while the French only sought to block flights to Paris that involved a plane that was smaller than the aircraft used for the U.S.-London portion of the flight route.[86]

A fair reading of *Air Service* suggests that the victim State—the United States—receives a measure of deference for its chosen countermeasures. On this view, it is sufficient for a countermeasure to reflect an approximation of the initial action's impact and the response needed to encourage compliance.[87] The tribunal's telling term, "approximation," only makes sense if the countermeasure *exceeds* the impact of the initial action since, by definition, a countermeasure that is clearly *less* impactful than the initial action would comply with proportionality.

[84] Zana v. Turkey, App. No. 18954/91, ¶ 51(ii) (1997) (ECtHR), http://hudoc.echr.coe.int/eng?i=001-58115 (noting that despite the protection of free speech, the Court upheld the criminal conviction of an official who used the phrase "national liberation movement" to describe a Kurdish group that Turkey had designated as a terrorist organization); Robert D. Sloane, *Human Rights for Hedgehogs?: Global Value Pluralism, International Law, and Some Reservations of the Fox*, 90 BOSTON UNIVERSITY LAW REVIEW 975, 983 (2010).

[85] *See* MICHAEL A. NEWTON & LARRY MAY, PROPORTIONALITY IN INTERNATIONAL LAW 183 (2014) (asserting that proportionality in countermeasures involves a "rough contextual approximation" based on the judgment of "policymakers acting in light of the information and assessments reasonably available to them to inform good-faith decision-making"); *see also id.* at 186 (describing proportionality in countermeasures as "prohibition against excesses rather than a requirement for equivalence").

[86] *Air Service Agreement*, *supra* note 5, at 443–44, ¶ 83.

[87] *Id.*

The *Air Service* tribunal may have taken this view because requiring a precise fit—not merely an approximation—between an initial violation and a countermeasure would have relegated victim States to responses that were less impactful than the initial action. Since a precise fit is hard to achieve, a victim State seeking to comply with international law under a precise-fit standard would need to set its response at a level *lower* than the initial action. In many situations, such a threadbare countermeasure would be manifestly inadequate to induce the responsible State to comply with its duties. Ensuring that a countermeasure is not a futile exercise thus requires a measure of deference for the victim State. Similar logic should govern in the cyber realm.

Even with an appropriate margin of appreciation, a suitably "commensurate" countermeasure should not interfere with an interest that is markedly more important than the interest that the initial action of the responsible State impaired.[88] Here, too, the *Air Service* decision is a useful guide. The U.S. countermeasure of threatening a halt to Air France's Paris-Los Angeles flights possessed "some degree of equivalence" with France's initial action in barring changes in the size of planes for the London-Paris leg of U.S.-France air routes.[89] Both the initial action by France and the U.S. countermeasure concerned the regulation of international commercial air travel. A U.S. response that disrupted internal French law enforcement, communications, or financial networks would arguably have targeted a French interest more important than the United States' interest in avoiding snarls in international air travel. That hypothetical U.S. countermeasure would thus have been disproportionate.

A disparity in the importance of interests affected may have also contributed to the International Court of Justice's (ICJ) decision in the *Gabčíkovo-Nagymaros Project* case.[90] In *Gabčíkovo-Nagymaros*, Czechoslovakia responded to Hungary's refusal after the fall of the Iron Curtain to implement a Warsaw Pact-era bilateral treaty on damming the Danube River. Czechoslovakia's countermeasure entailed unilateral action to "assume control" of the "shared resource" of this great river and divert the Danube's waters from Hungary.[91] That Czech action, if permitted to remain in effect,

[88] *See Draft Articles of State Responsibility*, *supra* note 80, at 135, cmt. ¶ 6.

[89] *Air Service Agreement*, *supra* note 5, at 443–44, ¶ 83.

[90] Gabčíkovo-Nagymaros Project (Hung./Slovk.), Judgment, 1997 I.C.J. Rep. 7 (Sept. 25).

[91] *Id.* at 56, ¶ 85.

would have deprived Hungary of its entitlement to an "equitable and reasonable share" of the Danube's natural resources.[92] The Czech fiscal interests affected by Hungary's refusal to honor the earlier agreement were far less important than the Hungarian riparian rights that the Czech countermeasure would have harmed.

On the other hand, an initial action's impact on a victim State's election integrity would rise to the foremost level of importance. Suppose, as the U.S. intelligence community has concluded, that Russia engaged in massive Internet-based efforts to sway U.S. voters during the 2016 election that relied heavily on deception, including fake Twitter and Facebook accounts.[93] Further assume that Russia sought to undermine the legitimacy of the 2018 U.S. congressional election results by using its Internet acolytes—human and machine—to spread false information about the outcome. Given the place of elections in democratic governance, interference of this type would impinge on a vitally important interest. The United States apparently responded to these Russian actions with a DDoS attack that put a key Russian government entity offline.[94] The reported U.S. action against a Russian agency believed to have been responsible for web-based election interference clearly did not exceed the importance of election integrity, which Russia had sought to undermine.

In addition to the importance of the interests that a countermeasure effects, a countermeasure should be temporary and reversible. The Draft Articles on State Responsibility make duration a criterion separate and distinct from proportionality, asserting that countermeasures are limited to acts or omissions that, (1) clash with the victim State's duties under international law, and, (2) operate "for the time being" in order to induce compliance by the responsible State.[95] In addition, the Draft Articles indicate that, where possible, countermeasures shall not irreversibly disrupt the status quo; rather, a victim State should take countermeasures in "a way . . . [that will] permit the resumption of performance" of the victim State's duties.[96] A

[92] *Id.*

[93] *See* Schmitt, *supra* note 26.

[94] *See* Ellen Nakashima, *U.S. Cyber Command Operation Disrupted Internet Access of Russian Troll Factory on Day of 2018 Midterms*, WASHINGTON POST, Feb. 27, 2019, https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html.

[95] *Draft Articles of State Responsibility*, *supra* note 80, art. 49(2).

[96] *Id.* art. 49(3).

countermeasure that fundamentally alters the relationship between the responsible and victim State will block such a "resumption of performance" of duties under international law, and thus will exceed a countermeasure's permissible dimensions.

Consider again the ICJ's decision in *Gabčíkovo-Nagymaros*. The water-diversion project that the Czech Republic had begun after Hungary's repudiation was far-reaching, with an impact on Hungary's riparian rights that would have been long lasting and virtually impossible to reverse. As the ICJ observed, a vital concern in Czechoslovakia's unilateral diversion of the Danube was the "continuing effects of the diversion . . . on the ecology" of Hungary's riparian region.[97] While the Court framed these effects as part of its finding that Czechoslovakia's response was not proportional, the Court's discussion of the Czechoslovakian response of "continuing effects" on Hungary's ecological systems suggests that duration and proportionality often go together.[98] Common sense indicates that "continuing effects" on sensitive ecological systems often will be lasting. Given the difficulty of reversing such effects, it makes sense to integrate duration into the proportionality calculus.

In the cyber realm, the durational limitation suggests that a DDoS attack would generally comply with proportionality in countermeasures. A DDoS attack does not destroy functionality or require replacing hardware or software, and hence remains below the use of force threshold. For the same reason, a DDoS attack is reversible; once it stops, the network returns to its prior level of functioning. Moreover, DDoS attacks do not destroy data; they just make data on a network—like the network itself—more difficult to access for some period of time.

However, the durational criterion suggests that classifying a DDoS attack as a proportionate countermeasure is both over- and under-inclusive. First, a DDoS attack has no built-in time limit. With a large enough botnet, an adversary can continue a DDoS attack for days, weeks, or months. At some point, a sustained DDoS attack becomes permanent, not merely temporary. Moreover, a sustained DDoS attack—or even one of more modest duration—can have effects that are irreversible and reasonably foreseeable as such. For example, a DDoS attack that has foreseeable collateral adverse effects on a financial network can generate opportunity costs that hurt investors or other stakeholders; individuals who lose the ability to trade stocks for a period of time can miss out on a chance to buy a stock that rapidly

[97] Gabčíkovo-Nagymaros Project, 1997 I.C.J. Rep. at 56, ¶ 85.

[98] *Id.* at 56–57, ¶ 87 (expressly disclaiming decision on issue of reversibility).

increases in value, sell a stock whose value precipitously declines, or engage in short-selling to profit from a decline in share price. In the countermeasure context, such effects may in the aggregate be disproportionate.

Moreover, just as a DDoS attack may be disproportionate, depending on the importance of the system affected, an attack on the functionality of an operating system is not necessarily disproportionate on the view taken in this article. The next Section explains that the cyber realm's dynamic nature permits inquiry on whether a prompt patch or other repairs will address the loss of functionality in a timely fashion.[99] Prompt patching may entail information from the victim State whose countermeasure is causing the loss in functionality. Suppose the State conducting the cyber operation provides information on restoring functionality directly to a State that has experienced the loss—for example, a neutral State that has experienced collateral damage—or indirectly to a company that uses the affected software or an international organization set up for the purpose of restoring functionality. In that event, the timeliness and comprehensiveness of the information provided would reduce the harm that figures in the proportionality calculus. In proportionality under the *jus in bello*, a State's willingness to provide a timely and comprehensive patch therefore relates to the quantum of harm to civilians that is reasonably expected by those planning a countermeasure or intrusion. Of course, even a timely and comprehensive patch may not prevent irreversible harm, such as opportunity costs that arise from the loss of functionality before the patch becomes effective. Those harms will continue to count toward proportionality.

In sum, proportionality should generally consider duration and reversibility together. Generally, a temporary countermeasure will best fit the proportionality rule. Countermeasures that take longer to reverse are more likely to be disproportionate. After all, the effects of many kinetic attacks, including bodily wounds and damage to property, may also be "reversible" given enough time, effort, money, and expertise. Actions causing such effects would exceed the use of force threshold, but the example above shows the folly of separating reversibility and duration.[100]

[99] A patch is a change in software that will address a flaw or "vulnerability" that made the software susceptible to hacking. *See* Tim Ridout, *Building a Comprehensive Strategy of Cyber Defense, Deterrence, and Resilience*, 40 FLETCHER FORUM OF WORLD AFFAIRS, Summer 2016, at 63, 70.

[100] I am indebted to Gary Brown for this point.

C. *The Duty to Take Feasible Precautions: An Express or Implicit Duty*

This article argues that whenever proportionality, in each of its guises, is applicable, the duty to take feasible precautions also applies, either expressly or implicitly. The rule of precautions is express in the *jus in bello*,[101] but also applies to the *jus ad bellum* and the law of countermeasures. As we shall see, the duty to take feasible precautions may either stand on its own, as an independent substantive duty layered on top of proportionality, or may be evidentiary in nature, demonstrating a State's *compliance* with the rule of proportionality. Both the substantive and evidentiary conceptions are important in the cyber domain because the need for speed often makes prompt action necessary, while also requiring feasible measures to mitigate the harm that speed could cause.

Under the rule of precautions in IHL, a State must take all "feasible" steps to reduce civilian harm.[102] A feasible step is one that is practicable, given resource constraints, technological limits, and tactical concerns, such as the importance of preserving certain means or instrumentalities of warfare (including weapons) for future engagements, and the disadvantage of disclosing certain advancements to adversaries or the world at large.[103] A feasible step is not one that is merely *possible*; requiring a State to implement all possible steps would unduly burden commanders, undermining the crucial value of military necessity.[104] But a definition of feasibility that imposed no duties on States would drain all meaning from the rule of precautions.

At the intersection of technology and the rule of precautions in attack, resource constraints and tactical concerns recede over time. As the mass production of any technology increases, it also becomes more widespread and less expensive. Moreover, knowledge of a once rare or closely-held technology typically proliferates, as, for example, the capacity to construct and deploy nuclear weapons increased from the time that the United States used nuclear weapons at the close of World War II to present. Decreased expense

[101] Additional Protocol I, *supra* note 4, art. 57(2)(a)(ii); Corn, *supra* note 6, at 459; Geoffrey Corn & James A. Schoettler Jr., *Targeting and Civilian Risk Mitigation: The Essential Role of Precautionary Measures*, 223 MILITARY LAW REVIEW 785, 837 (2015); Jean-Francois Queguiner, *Precautions under the Law Governing the Conduct of Hostilities*, 88 INTERNATIONAL REVIEW OF THE RED CROSS 793, 797 (2006).

[102] Additional Protocol I, *supra* note 4, art. 57(2)(a)(ii).

[103] *See* David A. Wallace & Shane R. Reeves, *Protecting Critical Infrastructure in Cyber Warfare: Is It Time for States to Reassert Themselves?*, 53 UC DAVIS LAW REVIEW 1607, 1635 (2020).

[104] *Cf.* Schmitt, *supra* note 61.

and increased production ease resource constraints and tactical concerns, making it more feasible to deploy formerly new technology.

The rise in technology that is evident in cyber and autonomy also makes precautions relevant in areas where they have not traditionally been salient, including countermeasures and the *jus ad bellum*. Technology highlights the need for speed—the importance of responding quickly to avoid greater damage or disadvantage and to increase the probability that a given measure by a victim State will effectively repel an incursion and persuade the responsible State to cease its offending conduct.[105] A victim State that is slow to respond encourages other States to violate international law, either with an armed attack that violates the *jus ad bellum* or with an action that violates the principle of sovereignty or constitutes an unlawful interference under the use of force threshold. However, the importance of speed in a response may also produce greater adverse impacts for the responsible State and for third party States.

Here, the rule of precautions has a substantive role to play, not only in the *jus in bello*, but also in the *jus ad bellum* and countermeasures. Suppose a State can feasibly deploy technology to craft a timely, effective countermeasure that is also more precise than other available responses. Given this assumption, this article argues that the rule of precautions applies and that as a result, States have a duty to deploy that more tailored technology.

Both the U.S. government and a spectrum of international law scholars have indicated support for a rule of precautions that would apply regarding the use of force, the conduct of hostilities, and countermeasures below the use of force threshold. For example, the U.S. Department of Defense has indicated that even below the use of force threshold, a cyber operation "should not be conducted in a way that unnecessarily causes inconvenience to civilians or neutral persons.[106] One distinguished commentator has criticized this statement by the U.S. Department of Defense as lacking adequate support or as merely stating a U.S. policy preference rather than articulating a binding legal requirement.[107] However, the U.S. Department of Defense's

[105] *See* Dan Saxon, *A Human Touch: Autonomous Weapons, Directive 3000.09, and the "Appropriate Levels of Human Judgment Over the Use of Force*," 15 GEORGETOWN JOURNAL OF INTERNATIONAL AFFAIRS, Summer/Fall 2014, at 100, 103–04.

[106] OFFICE OF THE GENERAL COUNSEL, U.S. DEPARTMENT OF DEFENSE, LAW OF WAR MANUAL § 16.5.2 (rev. ed., Dec. 2016).

[107] *See* Gary D. Brown, *Commentary on the Law of Cyber Operations and the DoD Law of War Manual*, *in* THE UNITED STATES DEPARTMENT OF DEFENSE LAW OF WAR MANUAL: COMMENTARY AND CRITIQUE 337, 346 (Michael A. Newton ed., 2018) (stating that military lawyers who rely on the DoD *Law of War Manual* "would be better served if the Manual

unqualified statement of a duty to avoid needless inconvenience to civilians through cyber operations is consistent with the substantive conception of precautions outlined here.[108]

The *Tallinn Manual 2.0*'s International Group of Experts seems to endorse such a role for precautions in countermeasures by citing the need to employ "considerable care" in crafting a proportionate countermeasure.[109] Indeed, the International Group of Experts suggest that prior to initiating countermeasures, a victim State must conduct a "full assessment" that includes "mapping the targeted system" and "reviewing relevant intelligence."[110] As was argued earlier in this Part and is discussed further in the next Part, taking such precautions does not necessarily lock a victim State into a rigid time sequence. A State can engage in such precautions *before* an attack or other action by a responsible State. Indeed, a prudent State would continually acquire cyber, signals, and human intelligence about its adversaries. The key point is that States have a duty to take such measures where feasible to temper the State's response or that response's effects. The emphasis on such steps suggests that under international law countermeasures include a precautionary element.[111]

Professor Michael Schmitt, the general editor of *Tallinn Manual 2.0*, recently outlined a comparable view of the importance of precautions. Discussing contexts at or below the use of force threshold, Schmitt argued that as a matter of policy, States should not engage in cyber incursions in which "[the] expected concrete negative effects on . . . the civilian population are excessive relative to the [anticipated] concrete benefit."[112] Although this advice adopts the language of proportionality, it also suggests a role for precautions.

made clear this [avoiding unnecessary inconvenience to civilians or neutrals] is a US policy rather than the law," which at most warranted placement in U.S. rules of engagement); *see also* Schmitt, *supra* note 55, at 349, 349 n.82 (describing the DoD *Law of War Manual* statement as addressing policy).

[108] The analysis in this article supplies additional analytical support for the U.S. Department of Defense position.

[109] TALLINN MANUAL 2.0, *supra* note 1, at 128.

[110] *Id.*

[111] Perhaps the International Group of Experts suggestion here largely pertains to the evidentiary conception—proving that the State engaging in countermeasures relies on best practices to facilitate compliance with the rule of proportionality. However, one can also read the *Tallinn Manual 2.0* analysis as recognizing that a substantive view of precautions is inherent in the requirement that countermeasures be proportionate.

[112] Schmitt, *supra* note 55, at 347.

To discern the role of precautions, suppose a State can reap a particular benefit with a cyber countermeasure at the cost of inconvenience to civilians at level X. Now suppose that the State can feasibly achieve the same benefit with a technological precaution that would reduce negative effects on civilians to one-half X. If the State decides to proceed *without* employing the feasible technological precaution—even though using the precaution would reap the same benefit—it is reasonable to view the difference between X and one-half X as "excessive." Professor Schmitt's description of the results of his balancing test supports this reading. For example, Professor Schmitt has noted that as a matter of policy a State should reject a cyber action that would yield significant civilian inconvenience when the expected benefit was "trifling."[113] Such a State decision, according to Professor Schmitt, would seem petty and mean-spirited.[114] Indeed, such a decision would not serve the criterion of military necessity that interacts with the principle of humanity to form IHL's crucial balance.[115] At least when the cost of employing a feasible precaution is *de minimis* because of economies of scale, a failure to employ that precaution would similarly fail Professor Schmitt's test.

Moreover, even if precautions do not have the freestanding substantive significance in countermeasures that they possess in IHL—imposing duties *beyond* proportionality when added safety steps are "feasible"—precautions do have an *evidentiary* significance. Here, a State that takes precautions before engaging in countermeasures can cite those precautions as evidence that its response is proportionate. For example, suppose a victim State's countermeasure includes collateral damage to the responsible State's systems or to neutral States that is more substantial in scale than the impact of the responsible State's initial action. As noted earlier, a State should receive a margin of appreciation that would cover modest increments beyond the initial action's effects. Suppose, however, that a margin of appreciation is only available when the victim State has demonstrated good faith or even reasonable care. In that event, the victim State's care in mapping the responsible State's system can constitute evidence that the victim State's actions causing the additional damage were not intentional, knowing, or even negligent. In this sense, the good faith and due care that a victim State shows through the taking of precautions is probative evidence of compliance with international law.

[113] *Id.* at 349.

[114] *Id.* (noting that such an incursion would "smack of mere maliciousness").

[115] *Id.*

D. Summary

This discussion has analyzed proportionality in the cyber domain in the contexts of the *jus ad bellum* and countermeasures. As noted in this Part, proportionality is both functional and substantive. Analyzing proportionality includes assessing whether the proposed countermeasure will, (1) elicit compliance with international law by the responsible State, and, (2) correspond with the importance, scale, and duration of the initial incursion. Under the view expressed here and supported by the *Air Service Agreement* arbitral award, a victim State has a margin of appreciation on the second criterion. Without that flexibility, a victim State facing rigid legal norms in an uncertain operational environment may choose a *more modest response* than the initial incursion. That restricted response would not effectively signal to the responsible State that the latter should cease its violation of international law. This Part has also argued that in addition to being an independent requirement under IHL, the need to take feasible precautions is inherent in both the *jus ad bellum* and countermeasures. In each case, precautions have a substantive dimension, imposing additional duties even when a State has satisfied proportionality, and evidentiary significance, demonstrating that the State has procedures in place that will promote its compliance with the proportionality rule.

IV. PRECAUTIONS IN THE USE OF AUTONOMOUS CYBERAGENTS

Now that we have discussed the test for proportionality and made the case for an inherent rule of feasible precautions in the *jus ad bellum* and countermeasures joining the express rule in the *jus in bello*, it is time to focus more specifically on the criteria guiding the rule of precautions. The use of autonomous agents in the cyber domain poses special challenges because of autonomous agents' brittleness, bias, and unintelligibility, as well as humans' tendency toward automation bias.[116] *Lex lata* has not yet caught up with the demands in this emerging arena. As such, the following discussion ventures into the venue of *lex ferenda*, although the discussion proceeds on the as-

[116] The U.S. Department of Defense has recognized the importance of these issues. *See* U.S. Defense Innovation Board, AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense: Supporting Document 31–33 (2019), https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF (discussing need to combat bias); *id.* at 33–38 (discussing the need to understand data inputs and review outputs for autonomous agents).

sumption that norms will move in this direction. By way of criteria for precautions, the article suggests four pillars: reconnaissance, coordination, repairs, and review. I address each in turn.

A. Reconnaissance and the Imperative of Espionage

While espionage and reconnaissance are mainstays of State behavior during peace and during war, this article goes further by arguing that intelligence collection, including espionage, is not merely permitted but *required* in the use of autonomous cyberagents. That requirement extends not merely to the *jus in bello*, which mandates consideration of "reasonably available" information in targeting decisions,[117] but also to the *jus ad bellum* and countermeasures. In the cyber realm, intelligence collection—which I refer to as reconnaissance—will often be virtual.[118] But on occasion, human aid to assist such efforts is necessary for their success—as in the case of human insertion of a thumb drive to introduce a worm for exfiltration of data.[119] In such situations, and where feasible, the approach taken here would require such human aid.

Virtual reconnaissance, supplemented as needed by human and signals methods of collection, is necessary to ensure that autonomous cyberagents comply with international law. Without the capacity to map an adversary's network and associated systems,[120] an autonomous cyberagent will be "flying blind," and will lack the ability to accurately target adversaries or avoid excessive collateral damage. As noted above, waiting until after an attack or other action has occurred will often hinder an effective response in each of the contexts examined here, including the *jus ad bellum* and countermeasures, as well as the *jus in bello*. Because of the need for speed, collecting cyber intelligence on potential adversaries before an attack will often be the only way to ensure that a response is both effective and tailored to avoid needless harm. Without that precaution, the brittleness and bias of autonomous

[117] *See* UNITED KINGDOM MINISTRY OF DEFENCE, JSP 383, THE JOINT SERVICE MANUAL OF THE LAW OF ARMED CONFLICT ¶ 5.3.4 (2004).

[118] TALLINN MANUAL 2.0, *supra* note 1, at 168. Militaries often refer to the gathering of information as encompassing intelligence, surveillance, and reconnaissance. *See* Michael N. Schmitt & Sean Watts, *The Decline of International Humanitarian Law* Opinio Juris *and the Law of Cyber Warfare*, 50 TEXAS INTERNATIONAL LAW JOURNAL 180, 210–11 (2015). Purely for ease of reference, this article uses the term "reconnaissance" to connote the full range of intelligence collection, including espionage.

[119] TALLINN MANUAL 2.0, *supra* note 1, at 171.

[120] *Id.* at 128.

agents will produce errors that both reduce the reasonably anticipated benefits of the attack or countermeasure and increase the harms that a reasonable decisionmaker would expect.

Under the substantive conception of precautions outlined in the previous Section, these concerns about foreseeably reduced benefits and increased harms dictate that when prior collection through reconnaissance—including espionage—is feasible, it is required. *Tallinn Manual 2.0* recognized that international law does not bar espionage, per se, including monitoring and exfiltrating data.[121] This Article goes further by suggesting that the duty to take feasible precautions inherent in the law of countermeasures includes a duty to undertake espionage, where the latter is feasible.

When a victim State used reconnaissance as well as the other steps suggested below—such as coordination, review, and repair—under an evidentiary view of precautions such measures would be presumptive evidence of proportionality, entitling a victim State to a margin of appreciation. On the other hand, suppose that reconnaissance is *not* feasible, and that in its absence a victim State cannot be reasonably certain that an autonomous cyberagent will be sufficiently precise to avoid excessive harm. Under the evidentiary view of precautions, use of the agent despite this concern would provide a basis to infer that the victim State had violated the rule of proportionality.

An example will be helpful. Suppose that Arcadia implants malware in various government networks of Pacifica. The malware has autonomous capabilities: it has been trained both to observe Pacifica's networks and react to particular inputs from those networks. Suppose further that Arcadia's autonomous malware receives inputs indicating that Pacifica has just commenced an attack on Arcadia's networks. Based on inputs about the operation of Pacifica's networks that Arcadia's malware has already gathered, Arcadia's autonomous cyberagent will be able to launch corresponding attacks on Pacifica's networks, echoing the scale, scope, duration, and importance of the attacks on Arcadia. Without the autonomous malware already in place, Arcadia would have had to "start from scratch" in both attributing and responding to the attack. Absent the autonomous cyberagent that Arcadia had already implanted in Pacifica's networks, Arcadia might have mistakenly attributed the attacks to another rival, Ruritania. By virtue of the malware it had previously implanted, Arcadia has the capacity to both correctly attribute the attacks to Pacifica and respond appropriately. If placing malware in

[121] *Id.* at 168–69.

Pacifica's networks is feasible, this would be a necessary precaution on Arcadia's part.[122]

To flip the hypothetical, suppose that such placement was not feasible and Arcadia could not reasonably find that a newly introduced autonomous agent would provide the benefit that Arcadia anticipated and avoid excessive foreseeable harm to both Pacifica and Ruritania. In that event, Arcadia's use of the new agent would presumptively violate the rule of proportionality, with a finding of culpability hinging on the actual damage to Pacifica and Ruritania's networks.

Use of feasible reconnaissance—including espionage—as a necessary precaution in both a substantive and evidentiary sense would be required even if the reconnaissance involved a physical intrusion into the territory of another State. For example, suppose that Arcadia could implant its autonomous cyberagent only through the use of a human agent who would insert a thumb-drive containing malware into a computer within Pacifica. *Tallinn Manual 2.0* did not reach a definitive conclusion on whether the use of a human agent on Pacifica territory would constitute an illegal violation of Pacifica's sovereignty.[123] In this respect, the *Tallinn Manual 2.0* International Group of Experts left open the possibility that Arcadia's action could be unlawful. The approach taken here takes a different path. It would clearly hold that espionage of this kind is entirely lawful, despite its location on Pacifica territory.[124]

On this view, espionage—as long as it merely involves exfiltration of data—is an international *public good*: an activity that disseminates knowledge about each State's capabilities and defenses and thus facilitates accurate cyber responses both at and below the use of force threshold. For the same reason, under the substantive and evidentiary conceptions of precautions advanced here, espionage with a territorial component is—where feasible—not only permitted, but *required*, at least if no other mode of reconnaissance will promote an autonomous cyberagent's reasonably accurate assessment of expected benefits and harms.

[122] If Pacifica detected Arcadia's malware and removed it, Arcadia would be required to deploy the "next generation" of malware, if that step were feasible.

[123] TALLINN MANUAL 2.0, *supra* note 1, at 171.

[124] *See* Ashley Deeks, *An International Legal Framework for Surveillance*, 55 VIRGINIA JOURNAL OF INTERNATIONAL LAW 291, 302 (2015) (suggesting that the permissibility of espionage informs the contours of sovereignty); *see also* David A. Wallace, Amy H. McCarthy & Mark Visger, *Peeling Back the Onion of Cyber Espionage After* Tallinn 2.0, 87 MARYLAND LAW REVIEW 205, 222–28 (2019) (discussing disagreements within the group of experts who convened to draft *Tallinn Manual 2.0*).

B. Coordination of Autonomous Methods

The brittleness and bias of autonomous agents, in addition to requiring increased reconnaissance, also mandates expanded coordination. By coordination, this article refers to the use of different autonomous methods simultaneously or in close succession to refine outputs.[125] The interaction of different modes acts as a check on agents' errors. A model with a particular strength or training in specific data can blunt the impact of arbitrary or biased outputs from other models with different strengths and training. Working together, coordinated models can perform more functions without gaps or mistakes.[126]

In cyber, coordination amounts to an autonomous "red team." Just as a "red team" makes better human decisions by posing objections and presenting alternatives, the autonomous equivalent lowers the risk of false positives while ensuring that attribution is precise.[127] The autonomous use of coordination aids greatly in both detecting anomalies and misuse in networks that may signal a cyber intrusion.

Many autonomous models build coordination into their approach, including two or more methods. For example, ensemble learning simultaneously applies a suite of hypotheses about inputs.[128] Each "learner"—that is, each discrete machine that analyzes a given set of inputs—generates a "weak" hypothesis, which operators define as a hypothesis about data that

[125] *See* SCHARRE, *supra* note 1, at 20–21; STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL (2019) (discussing the importance of checks on the outputs of any single autonomous learner).

[126] *See* RUSSELL & NORVIG, *supra* note 29, at 1005; DOMINGOS, *supra* note 29, at 238.

[127] Anna L. Buczak & Ethan Guven, *A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection*, 18 IEEE COMMUNICATIONS SURVEYS AND TUTORIALS 1153, 1162–70 (2016), https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7307098; *see generally* Mark Raymond, *Engaging Security and Intelligence Practitioners in the Emerging Cyber Regime Complex*, 1 CYBER DEFENSE REVIEW, Fall 2016, at 81, 92 (discussing human red-teaming in the cyber arena); U.S. CYBERSPACE SOLARIUM COMMISSION, *supra* note 21, at 22 (discussing red-teaming of preliminary policy proposals to identify their weaknesses). On issues of attribution in cyber incursions, see DENNIS BROEDERS, ELS DE BUSSER & PATRYK PAWLAK, THE HAGUE PROGRAM FOR CYBER NORMS, THREE TALES OF ATTRIBUTION IN CYBERSPACE: CRIMINAL LAW, INTERNATIONAL LAW AND POLICY DEBATES 7–8 (2020), https://www.universiteitleiden.nl/en/research/research-output/governance-and-global-affairs/three-tales-of-attribution-in-cyberspace.-criminal-law-international-law-and-policy-debates; Nicholas Tsagourias, *Cyber Attacks, Self-Defence and the Problem of Attribution*, 17 JOURNAL OF CONFLICT AND SECURITY LAW 229 (2012).

[128] Buczak & Guven, *supra* note 127, at 1164–65.

is imperfect, but performs better than a random prediction. Ensemble learning then combines each of the better-than-random predictions to generate a more robust hypothesis. That robust hypothesis will yield outputs that avoid the arbitrary results fostered by any one approach's tendency toward brittleness and bias. In assessing and attributing cyber intrusions, for example, an ensemble learning method or other approach demonstrating coordination would separate out false positives. As currently used in spam detection methods, a coordinated agent would distinguish phishing attempts from bona fide emails as well cyber initiatives from friendly or neutral States from malicious incursions by adversaries.

The coordination factor described here is not prescriptive regarding a *particular* autonomous methodology—the ensemble learning approach described above is merely an illustration. Coordination's core is an autonomous capability to conduct different inquiries of inputs simultaneously or in tight succession, to test preliminary hypotheses rapidly, and to weed out the effects of brittleness and bias. A State that has deployed an autonomous cyberagent with this coordination capability has checked another box in the precautionary matrix.

As an example, suppose that in our implanted malware hypothetical, Arcadia had planted malware in Pacifica's government software, but Pacifica—suspecting that Arcadia or other States had done this—had taken steps to throw them off the track. To accomplish this, Pacifica had deployed "adversarial examples" in visual representations of government departments as well as in code itself—changing digits in unused code spaces for the particular purpose of deceiving autonomous cyberagents from other States.[129] Further suppose that Pacifica then launched a cyber intrusion against Arcadia below the use of force threshold. If the adversarial example was effective, one Arcadian autonomous cyberagent might not be able to discern the difference between a network used by a Pacifica government agency and a civilian network.

[129] *See* Rouhani et al., *supra* note 37, at 34–35. Use of adversarial examples to muddy an attacker's ability to distinguish between military and civilian objects might violate a defender's duty to take precautions under IHL. *See* Additional Protocol I, *supra* note 4, art. 58(c); Samuel Estreicher, *Privileging Asymmetric Warfare? Defender Duties under International Humanitarian Law*, 11 CHICAGO JOURNAL OF INTERNATIONAL LAW 425, 432–36 (2011); Eric Talbot Jensen, *Precautions against the Effects of Attacks in Urban Areas*, 98 INTERNATIONAL REVIEW OF THE RED CROSS 147, 156–57 (2016) (explaining the defender's obligation to take feasible precautions to lower the risk of harm to civilians on territory within its control).

To ensure a proportional response, Arcadia would have to deploy ensemble learning or another coordinated model.[130] Pacifica's adversarial examples might well fool Arcadia if the latter State had only used one autonomous method. In contrast, using a coordinated model incorporating two or more disparate methods would greatly increase the accuracy and reliability of Arcadia's targeting and reduce the likelihood that Pacifica's use of adversarial examples would cloak its responsible network from an Arcadian countermeasure.[131] Fooling one method is certainly possible, but fooling two or more methods is far more challenging, particularly if one of those methods is trained to spot changes in the cyber topography that could be evidence of adversarial deployment. Indeed, Pacifica might be less willing to try its luck with an initial incursion, thus promoting greater compliance with international law.

Coordination like this is nothing new in a State or commander's lexicon. Commanders regularly use a range of inputs from intelligence, surveillance, and reconnaissance, and strive to weigh disparate inputs in a balanced fashion to reduce the chance of reliance on a single flawed source. Redundancy is also a common feature of automotive and aircraft software, weapons systems, and other advanced technology.[132] The coordination criterion merely builds on this foundation.

In the Boeing 737 Max incidents, the operation of a single sensor exaggerated the risk of a stall and thus triggered a downward plunge in the aircraft's nose that resulted in two catastrophic crashes.[133] Additional sensors would have more readily detected ambiguous information, suggesting that a stall was not imminent and that automatic depression of the aircraft's nose was not necessary. That redundancy in autonomous systems is one way that victim States can properly gauge connections between systems in States responsible for initial actions, thus ensuring that countermeasures minimize harm to unrelated systems in the responsible State.

One objection to coordination might be that coordinating different methods will consume more time, hindering an efficient response. But that

[130] Rouhani et al., *supra* note 37, at 34–35.

[131] *Id.*

[132] *See* Andre Kohn, Rolf Schneider, Antonio Vilela, Udo Dannebaum & Andreas Herkersdorf, *Markov Chain-based Reliability Analysis for Automotive Fail-Operational Systems*, 5 SAE INTERNATIONAL JOURNAL OF TRANSPORTATION SAFETY 30, 32 (2017). Fail-safe capabilities are a common feature of advanced systems. These features minimize risk in the event of malfunction.

[133] Hamby, *supra* note 53.

is not necessarily true. Autonomous agents excel at implementing tasks quickly, and Moore's Law suggests that the speed of applications will increase exponentially.[134] Any decrease in speed is likely to be minor and temporary, with timely response becoming increasingly frequent as engineers work out any initial glitches. The accuracy and reliability of coordination will also improve. Over time, rapid coordination will become increasingly feasible, leaving States that fail to take this course as outliers.

C. Repairs: A Patch in Time

As another modification that is appropriate for the cyber domain, the approach taken in this article requires—where feasible—that a State assist in repairs of collateral damage caused by autonomous cyberagents. In the cyber arena, a patch may often remedy damage quickly, in contrast with the time-consuming physical repairs that may be required for the effects of kinetic attacks. Suppose that a victim State's response has caused collateral harm to third-party States or unrelated or civilian networks in a responsible State. If the victim State can feasibly provide a timely and effective patch, the approach to precautions taken here would require that action. In addition, if the collateral harm fell below the use of force threshold, the law of countermeasures would require that—again, where feasible—the victim State provide a patch as part of its duty to ensure that the effects of countermeasures are both temporary and reversible.

In an armed conflict or even peacetime cyber exchanges with adversary States, sharing patches may be more difficult. For example, the United States may be wary of sharing software patches with adversaries such as Russia and China. In these situations, an international organization might be needed as an intermediary. Of course, IHL already uses trusted intermediaries for matters such as providing aid to civilians in war zones. Organizations analogous to the International Committee of the Red Cross or Doctors Without Borders could be established to act as clearinghouses for patching information. While sharing information may still not be practicable in such situations, the

[134] Lee Bell, *What is Moore's Law? WIRED Explains the Theory that Defined the Tech Industry*, WIRED, Aug. 28, 2016, https://www.wired.co.uk/article/wired-explains-moores-law (noting that according to the theory, originally expounded in 1965, that because of decreasing transistor size, the number of transistors per square inch of chip would double approximately every twelve months (now every twenty-four months)).

evanescence of vulnerabilities once used should encourage more information sharing, much as States share information with international health and humanitarian groups.[135]

As an example, suppose that Arcadia has used malware as part of a cyber countermeasure responding to an intrusion by Pacifica, but in the process impaired the functionality of software in Ruritania. Arcadia promptly acknowledged responsibility for the harm to Ruritania, and provided a patch that restored the functionality of adversely affected operating systems. Assuming that Arcadia incurred no substantial costs through this action that might have reduced its feasibility, the approach to precautions taken in this article would require that Arcadia provide the patch to Ruritania. Furthermore, while prompt provision of an effective patch would not completely remove the harm to Ruritania from the proportionality calculus applicable to countermeasures, it would reduce the quantum of harm used in this calculus.

D. Review: Unpacking the Unintelligible

In assessing how reconnaissance, coordination, and repairs have performed, review is essential. In IHL, review is part of a State's duty to exhibit "constant care" in reducing needless harm to civilians.[136] Proportionality in the *jus ad bellum* and countermeasures also requires review. A State that has engaged in a methodical review of past operations inspires trust that it will learn the right lessons from previous mistakes. That review should be independent to avoid the groupthink that can undermine neutral evaluation. Moreover, review in the autonomous cyber context depends on a State ensuring that its agents' outputs are sufficiently explainable to facilitate review.

In the *jus in bello*, review of a weapon starts prior to deployment with Article 36 of Additional Protocol I, which requires a finding that a weapon is not inherently indiscriminate.[137] Article 36 reviews have a low threshold: a

[135] Of the steps suggested here, the provision of repairs is both the least practicable and the greatest departure from the *lex lata*. It may be useful to consider State commitment to the three other steps outlined here—reconnaissance, coordination, and review—as an alternative approach that would still yield a margin of appreciation.

[136] *See* Additional Protocol I, *supra* note 4, art. 57(1).

[137] *Id.* art. 36; *see also* WILLIAM H. BOOTHBY, WEAPONS AND THE LAW OF ARMED CONFLICT 347–48 (2d ed. 2016); Michael W. Meier, *Lethal Autonomous Weapons Systems (LAWS): Conducting a Comprehensive Weapons Review*, 30 TEMPLE INTERNATIONAL AND COMPARATIVE LAW JOURNAL 119, 124–26 (2016). Even if an autonomous cyberagent passes a weapons review, designers will need to validate its use for particular purposes. *Cf.* Margaret

State need only find that some use of a weapon is consistent with IHL. For example, if a State can show that in a particular context, it can use a weapon to target an adversary's force, that weapon has met the requirements of Article 36. Review under Article 36 is vital where this duty applies, but it is more limited than the concept of review advanced here. First, a State's use of cyber may not be a weapon in the Article 36 sense of the term.[138] Second, Article 36 does not apply to countermeasures or other actions taken outside armed conflicts. Third, review here stems from the concept of after-action review in IHL.

After-action review in IHL and provisions for review under international human rights law (IHRL)[139] are more expansive in scope than pre-deployment Article 36 review. Because the combination of cyber and autonomy is so new, a review should be systemic, not merely focused on a specific incident. A State investigating an alleged war crime by one of its service members has no duty to consider whether it should forego the use of humans in future military engagements. The use of humans is sufficiently well established to render any such inquiry unnecessary. In contrast, depending on the seriousness of the outcomes, the novel technology of autonomous cyberagents may require a more searching review of the appropriateness of their deployment.

Such reviews entail a more robust form of independence than the fact-specific detachment required under customary IHL.[140] Under the *lex lata*, an investigation of alleged war crimes is sufficiently independent if it does not suffer from command influence that skews the investigation's analysis and conclusions. However, IHRL has been moving toward a more robust conception that requires greater structural independence from the chain of command.[141]

Under the approach taken in this article, a more robust structural approach would be required in IHL—recognizing the move in that direction in State practice—and in the *jus ad bellum* and countermeasures. Moreover, as a functional matter, the novelty and complexity of autonomous cyber-

Hu, *Small Data Surveillance v. Big Data Cybersurveillance*, 42 PEPPERDINE LAW REVIEW 773, 812–16 (2015) (urging the use of a rigorous test to validate machine learning models).

[138] *See* Jeffrey T. Biller & Michael N. Schmitt, *Classification of Cyber Capabilities and Operations as Weapons, Means, or Methods of Warfare*, 95 INTERNATIONAL LAW STUDIES 179 (2019).

[139] Michael N. Schmitt, *Investigating Violations of International Law in Armed Conflict*, 2 HARVARD NATIONAL SECURITY JOURNAL 31, 80 (2011).

[140] *Id.* at 50–51.

[141] *Id.* at 49–51.

agents counsel for such robust safeguards. In particular, while military commanders may well have exceptional expertise at their disposal, mobilizing expertise for reviews of agents' performance may require access to civilian developers and engineers. Indeed, civilian expert input may be necessary to ensure that a review is *effective*, which both IHL and IHRL also require.[142] Reviewers need a working knowledge of a technology to competently analyze where an action may have gone wrong. Without such analysis, an investigation will lack the effectiveness that international law demands.[143] On the other hand, a record of careful review would be one aspect of persuasive *evidence* that a State had complied with international law.

Reviews must include efforts to explain the outputs of autonomous agents. As noted earlier, explainability is a challenge for certain forms of artificial intelligence. In particular, neural networks generate outputs that are difficult to explain through conventional verbal means, since the layers that contribute to neural networks' accuracy sift through so many variables. To accommodate this concern, reviewers will need to define explainability more broadly, while ensuring that reviews are rigorous.

A broader conception of explainability would be multi-pronged and entail the machine coordination discussed above. One approach would be functional, seeking to clearly identify the interaction between inputs and outputs. This approach would use counterfactuals by varying a range of inputs and studying the outputs that the agent produced.[144] In coordination with other autonomous agents, such reviews would seek to isolate the mix of variables that yielded a decision that violated international law. If an agent weighted one variable too heavily, designers would retrain the agent using revised training data that diminished this variable's role. In addition, designers are working on agents that can express a neural network's outputs

[142] *Id.* at 80. Here, too, Professor Schmitt suggests that context should be determinative, and that the "complexity of the matter" should drive the scope, structure, and operation of the particular investigation. The approach taken in this article would vary from Professor Schmitt's approach only in making the *ex ante* decision that the complexity of autonomous cyberagents requires a greater measure of independence, at least until "proof of concept" has taken hold and a State has resolved systemic issues.

[143] Not every incident requires review. To require investigation, an incident must entail at least a colorable violation of international law. *See id.* at 79.

[144] Lipton, *supra* note 13, at 6; Lehr & Ohm, *supra* note 29, at 692; Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARVARD JOURNAL OF LAW AND TECHNOLOGY 841, 881–83 (2018). For further discussion, see *supra* note 49 and accompanying text.

through other more intelligible methods, such as decision trees that graphically depict factors that contribute to a decision. Once designers have created a decision tree that matches a neural network's outputs, the designers can "prune" the tree, cutting off irrelevant branches. Pruning can also yield a better mix of training data.

A workable conception of review should recognize that there are many ways of enhancing explainability and addressing errors. Moreover, a State should be able to show that it is continually working on more effective means for addressing this concern. Commitment to a reasonable framework of review is more productive than prescribing or prohibiting a particular technology.

In our malware hypothetical, a review might be required to determine the cause of mistakes and seek to correct tactics, techniques, and procedures in the future. For example, suppose that Arcadia had used malware embedded in Pacifica's networks to respond to a Pacifica incursion, but that malware had targeted civilian networks in a fashion that manifestly violated the *jus in bello*, *jus ad bellum*, or the rule of proportionality in countermeasures. Arcadia would be required to conduct a review to determine the cause of its mistake and discern means to avoid comparable mistakes in the future. Conducting that review would entail the capacity to discern *why* the agent made a mistake. For example, designers reviewing the agent's performance could seek to reverse-engineer that performance with counterfactuals to determine what inputs or architecture would have to change to secure a different result.

Upon review, designers could determine that they needed to use more elaborate coordination between autonomous learners to detect potential errors and modify the agent's outputs before they created harm. Under the approach taken here, designers would then have to implement the findings of their review. That dedication to review would diminish brittleness, bias, and unintelligibility and facilitate continual improvement in compliance with international law.

V. Conclusion

In the cyber realm, where the need for speed is paramount, observing proportionality is crucial. Human designers and operators lack the agility to respond to ever-mounting cyber incursions. Autonomous cyberagents can address that need.

In the *jus ad bellum*, *jus in bello*, and the law of countermeasures, proportionality plays an important role in reducing harm and the risk of escalation.

However, the amorphous character of proportionality makes it difficult to implement across each of the legal arenas described above. Attributes of autonomy also hinder that mission. Along with their extraordinary speed and analytical prowess, autonomous agents have notable flaws, including brittleness, bias, and unintelligibility. Beset by automation bias, human designers and operators struggle to accept and address these flaws.

Unduly burdening victim States is no answer to autonomy's deficits. In decisions about the use of force, the conduct of armed conflict, and the launching of countermeasures, overly onerous restrictions will force victim States to cede the initiative to first movers who violate international law in search of an advantage. States will reject any legal duty that yields this perverse result. A balance that encompasses the need for speed in victim State responses while ensuring that those responses remain within reasonable bounds is both desirable and necessary.

The approach taken in this article seeks to accomplish that goal. It confers a margin of appreciation on victim States' responses. However, that margin of appreciation requires victim States to observe feasible precautions. Those precautions have both independent substantive significance as a component of proportionality and evidentiary value as proof of a victim State's compliance with international law. Necessary precautions are reconnaissance, coordination, repair, and review. Fulfilling those conditions will allow victim States to wrest the initiative from the offending States while keeping their own responses in check. That balance will preserve stability in the cyber domain and the international order, while also complying with international law.