Garbage In, Garbage Out: Data Poisoning Attacks and their Legal Implications

Lieutenant Colonel Mark Visger

I. Introduction

With each new development in the cyber era, a corresponding attempt to "hack" the new technology has taken place. For example, in the case of self-driving cars, experts have demonstrated how such a car can be hacked, with the hacker able to assume control of the vehicle.[1] With the advent of Artificial Intelligence (AI) powered by machine learning and the prospect of deploying such tools into combat systems, the obvious question is raised—can AI be hacked?

The process of hacking usually requires deep technological understanding of the structure and mechanics of the underlying technological system. The hacker needs to dive deep "under the hood," so to speak, to find flaws and bugs to exploit. In the case of artificial intelligence/machine learning, hackers seek to do the same thing. In the case of AI, and more specifically machine learning, when one looks under the hood, one finds big data. Many have said that "data is the new oil,"[2] in part because data is the underlying substance that is able to power analytic tools such as machine learning and predictive analysis.[3] For purposes of this chapter, big data will be defined as "a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications."[4] These analytic tools distill this big data into actionable information—such as the recommendations that are now ubiquitous on cyber commerce websites.[5] These machine learning systems process large amounts of data as part of the "learning" process—making connections, discerning inferences, and ultimately drawing conclusions in a process that looks somewhat like what we humans call "learning."[6] As a result, if one can hack the big data that powers machine learning, one can potentially also hack the machine learning process itself.

To help visualize this process, let's consider a somewhat absurd hypothetical. Suppose an archetypical villain from a James Bond movie develops a machine learning model to identify cats from real-time video feeds. This information is fed to a loitering weapon, which we'll call the de-catenator, which targets the cats for destruction. In order to develop an application that identifies cats in real time, the villain would need to create a machine learning model which is trained to identify cats by "watching" thousands of hours of cat videos. In one type of machine learning, the cats in these videos would be labeled so that the machine learning can discern the relevant features on the labeled cats and "learn" how to identify cats independently. When the learning process is complete, the system would

[1] Eric A. Taub, *Carmakers Strive to Stay Ahead of Hackers*, N.Y. Times (Mar. 18, 2021), https://www.nytimes.com/2021/03/18/business/hacking-cars-cybersecurity.html.

[2] This quote is widely attributed to Clive Humby. Charles Arthur, *Tech Giants may be Huge, but Nothing Matches Big Data*, The Guardian (Aug. 23, 2013, 3:21 PM), https://www.theguardian.com/technology/2013/aug/23/tech-giants-data. The purpose of the analogy was two-fold, to state its great value in driving these new processes but also to state the need for refinement, in order for big data to be valuable. *Id.*

[3] Bridget Botelho & Stephen J. Bigelow*, Big Data,* TechTarget.com, https://searchdatamanagement.techtarget.com/definition/big-data (last visited Oct. 29, 2021).

[4] *Id.*

[5] *Id.*

[6] Anthony D. Joseph et al., Adversarial Machine Learning 20-21 (2019).

be deployed operationally, able to identify cats from video feeds in real time and funneling this information to the loitering weapon, threatening the feline species with imminent destruction. Now suppose that a hacker, who loves cats but hates dogs, accesses the training data and swaps all of the "cat" and "dog" labels. This hack "poisons" the data, and it is now the dogs who will meet their demise.

It is not hard to make the leap from cats and dogs to soldiers and combatants. In fact, the short video *Slaughterbots* claims that such technologies are imminently available and presents a dystopian scenario where such weapons are deployed against civilian human rights activists using AI-enabled facial recognition technology.[7] In fact, a UN report states that lethal autonomous drones have been used on the battlefield in Libya.[8] One can easily visualize AI making its way onto the battlefield, one possible example being the AI-enabled targeting system described in the Scenario chapter. Because one would expect military machine learning systems themselves to be hardened with protections against cyber attacks, direct hacks of the machine learning algorithms are likely to be impossible. Instead, much like switching the cat and dog labels, attacking the large amounts of data being fed to the machine learning algorithms so that the targeting system's performance is degraded or completely inaccurate will be a likely way to counter them.

While many are thinking about the legal frameworks for using AI in combat, such as to support targeting, the scholarly literature is relatively devoid of discussion about the legal framework to apply to attacks on the *data* upon which the AI system relies—either during the training phase or when operationally deployed. This is a significant oversight. If AI artificial agents employed on the battlefield have the decisive effect that experts are predicting, then enemy forces will accordingly train their fires—both kinetic and cyber—to attempt to neutralize the AI agent. Further, cyber "virtual fires" on an AI system's data is likely to have physical impacts, such as weapons-systems malfunctions which result in civilian casualties. Unfortunately, it is at best unclear how current legal frameworks apply to attacks on "mere" data. The dominant approach to this question is quite limited—to an analysis of whether or not an attack on data has impact in the physical environment.[9] As data becomes much more central to operations powered by AI, the legal approaches and frameworks must similarly account for the significant changes in how data is used to power AI.

The prospect of data poisoning attacks creates potential implications for the application of the Law of Armed Conflict (LOAC) that should be addressed before such attacks become a reality.[10] The nature of AI and the potential for attacks on the big data underlying the AI result in novel issues that have not yet been addressed. Three primary issues are immediately apparent. First, there is the very real prospect that data poisoning activities will take place before armed conflict even begins, while the machine learning models are being trained. Is such pre-conflict activity governed by the LOAC? If so,

[7] Stop Autonomous Weapons, *Slaughterbots*, YouTube (Nov. 12, 2017), https://www.youtube.com/watch?v=9CO6M2HsoIA.

[8] Final Rep. of the Panel of Experts on Libya (2021), transmitted by Letter dated 8 March 2021 to the President of the Security Council, ¶ 63, U.N. D s/2021/229 (Mar. 8, 2021).

[9] TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS, 416 (Michael N. Schmitt & Liis Vihul eds., 2017) [hereinafter TALLINN 2.0].

[10] In order to understand how data poisoning might be used in armed conflict, it will be helpful for the reader to review the Scenario chapter at the outset of this volume. This chapter will draw from the scenario which describes data poisoning and data evasion activities by the fictional country Outlandia against its enemy Newtropia, who had deployed AI targeting, called the Newtropian AI Targeting System (hereinafter NAITS).

how? Second, the effects of such data poisoning efforts are highly uncertain and attenuated. How can a commander apply targeting concepts like proportionality and military necessity in operations targeting big data via data poisoning in such an uncertain environment? Third, do such operations create the potential for perfidy, particularly in situations where the data poisoning results in conclusions that enemy forces are in fact protected persons?

A new academic sub-discipline has been created to catalog the methods by which machine learning can be attacked and secured against attack—Adversarial Machine Learning.[11] Academic experts in this field have crafted a draft taxonomy which catalogues the technical challenges in adversarial machine learning.[12] The most likely attack vectors will be against the data. Similarly, legal experts must begin systematically considering the issues raised and begin developing legal frameworks for the deployment of adversarial machine learning in the context of armed conflict. This chapter will address the LOAC implications of data poisoning. It will begin with an overview of data poisoning and data evasion, along with potential military applications. It will then proceed to examine the three legal issues that use of these techniques will likely raise: (1) the applicability of the LOAC to pre-conflict data poisoning, (2) navigating the highly uncertain and attenuated effects of data poisoning during armed conflict, and (3) whether certain applications raise perfidy concerns.

II. What is Adversarial Machine Learning? Technical Overview and Legal Roadmap

A. Introduction to Adversarial Machine Learning.

Before delving into the legal analysis, it is necessary to provide a more detailed description of how big data powers processes such as machine learning and how data poisoning can corrupt machine learning. Artificial intelligence is defined as "computer systems able to perform tasks that normally require human intelligence."[13] Machine learning frequently powers this AI using computer components to "learn from data to perform such tasks."[14] Of course, AI standing alone is not all that useful. Instead, the application of AI to specific "use-cases" is where the technology shows its value. In fact, AI can be considered as building blocks for a large variety of applications. Another valid comparison for AI is electricity—useful not in itself, but instead through the various ways that the electric current can be harnessed when there is general and widespread availability.[15] In the military context, the most high-profile (and controversial) use-case would be in AI algorithms that engaged and executed targeting decisions with little to no human involvement.[16]

[11] *See generally*, JOSEPH ET AL., *supra* note 6.

[12] NAT'L INST. OF STANDARDS AND TECHNOLOGY, DRAFT NAT'L INST. OF STANDARDS & TECHNOLOGY INTERAGENCY OR INTERNAL REPORT 8269: A TAXONOMY AND TERMINOLOGY OF ADVERSARIAL MACHINE LEARNING (2019) (*hereinafter* Draft NISTIR 8269).

[13] *Id.* at 1.

[14] *Id.*

[15] Shana Lynch, *Andrew Ng: Why AI is the New Electricity*, Insights by Stanford Business, (Mar. 11, 2017), https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity

[16] *See* HUMAN RIGHTS WATCH, LOSING HUMANITY: THE CASE AGAINST KILLER ROBOTS (2012), *available at* https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf [hereinafter LOSING HUMANITY].

As with any new computer process, adversaries or other hostile actors will seek to "hack" this artificial intelligence.[17] The field of "adversarial machine learning" has emerged, predicated on the fact that AI/machine learning systems can and will be attacked, studying the ways in which attacks may be carried out and how systems may be hardened against such attacks.[18] This section will utilize the draft taxonomy published by the National Institute of Standards and Technology, Draft NISTIR 8269,[19] in order to classify and define the various types of attacks that can be used to poison the data powering the machine learning systems.

Generally, machine learning begins with a system processing large amounts of data in order to build a prediction model (or algorithm) able to perform a desired task independently.[20] This training phase is defined as taking place as the machine learning analyzes the training data to develop a model which can be applied in the real-world.[21] The example cited above describes an instance of developing AI to identify cats: in one method, the system processes many photographs or videos where cats are labelled.[22] From this training data, the system can develop connections and inferences as to what features define a "cat." Data scientists term this learning process the testing, or inference phase. Once the testing phase is complete, the AI is deployed to conduct the task for which it was trained (this is termed the "operational phase"). Going back to our cat example, in the operational phase the AI will use the algorithm created in the testing phase and will monitor real-time video feeds to identify cats and feed the information to the de-catenator. The NISTIR 8269 framework distinguishes between attacks on data in these two phases, utilizing the term "data poisoning" for attacks on data taking place during the training phase and the term "evasion attacks" for attacks occurring during the operational or testing/inference phase.[23]

Data poisoning is further broken down into attacks on the data and attacks on the actual machine learning algorithm, called "logic corruption."[24] This latter attack is very effective (albeit difficult to accomplish) but essentially gives the attacker control over the AI system and its outputs. (Going back to our cat example, in a "logic corruption attack," the dog-hater would access the actual algorithm that had been developed to identify cats during the learning process and re-write the algorithm to instead identify dogs). As military AI systems are likely to be hardened against attack with significant cybersecurity measures, a direct attack on the algorithm along these lines is highly unlikely and will not be considered in this chapter.

Attacks on the data are much more feasible, and can take a number of different forms (with overly simplified examples from our cat scenario included in parentheses): (1) "data injection," where the adversary inputs additional data into the machine learning system in order to manipulate decision

[17] This is already being done in the context of machine-learning based cyber security use cases. Ilja Moisejevs, *Poisoning Attacks on Machine Learning*, Towards Data Science, (July 14, 2019), https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db.

[18] JOSEPH *ET AL.*, *supra* note 6, at 3.

[19] Draft NISTIR 8269, *supra* note 12.

[20] TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING 2 (2nd ed., 2017).

[21] Draft NISTIR 8269, *supra* note 12, at 2-3.

[22] Liat Clark, *Google's Artificial Brain Learns to Find Cat Videos*, WIRED (JUN. 26, 2012, 11:15 am), https://www.wired.com/2012/06/google-x-neural-network/. Note that this article details a technique by which the artificial intelligence is able to learn to complete this task without labels.

[23] Draft NISTIR 8269, *supra* note 12, at 2-3.

[24] *Id.* at 6-7.

boundaries and outcomes (here, the dog-hater would add additional training videos featuring cats that look like dogs); (2) "data manipulation," where existing data in the system is altered to manipulate decision boundaries and outcomes (here, the dog hater manipulates the data residing in the training set, perhaps by adding random data to the training videos that causes dogs to be identified as cats); (3) "label manipulation," where the attacker is able to modify the output labels produced by the algorithm as it is processed through the machine learning system (as described in the introduction, this would involve switching the dog and cat labels); and (4) "indirect poisoning," where the attacker modifies the data as it transits between the database and the machine learning system (accomplished, for example, by a wiretap that can alter the data in the cat videos as they travel from the training database to the machine learning system).[25]

Evasion attacks are similar to the data attacks outlined in the training phase, but they take place once the system is operational. These attacks seek to utilize "inputs that are able to evade proper output classification by the model."[26] In such attacks, the attacker attempts to identify "a small input perturbation [e.g., a manipulation or change to the data] that causes a large change in the loss function and results in output misclassification."[27] Evasion can occur through one of two methods—through accessing and altering the digital data before it arrives at the algorithm for processing (much like the indirect poisoning described above, except that this poisoning takes place after the machine learning system is operational) or through adding items or images that will cause the artificial intelligence to produce an incorrect output.[28] Examples of adding "digital camouflage" to confuse machine learning systems are seen in media reports of digitally-patterned shirts that cause a machine learning agent to fail to identify the wearer as a person (in effect rendering them invisible to the artificial intelligence).[29] In another example, researchers added small, innocuous pieces of tape to a traditional American stop sign, which caused the AI agent to classify the stop sign as a speed limit sign instead (with the potential for devastating consequences).[30]

One additional consideration from the Draft NISTIR 8269 that will also impact the LOAC analysis is the level of knowledge that an adversary has of the machine learning system. It makes intuitive sense that, the more an adversary knows about the functioning of an AI/machine learning system, the easier it will be to identify attack vectors through techniques such as data poisoning or evasion attacks. Additionally, the attacker with greater knowledge will have higher confidence levels in the outcomes of those attacks. The Draft NISTIR 8269 takes this level of adversary knowledge into account and classifies attacks accordingly: white box, gray box, and black box. In a white box scenario, the adversary has full knowledge of the machine learning system and data used to train it.[31] A white box situation would be unlikely for a hardened, combat-ready AI system, and a black box or gray box situation would be much more likely. In black box situations, an adversary has no knowledge except for the possibility of being able to observe input-output matches in real time (e.g., observing the operations of the AI and drawing

[25] *Id.*
[26] *Id.* at 7.
[27] *Id.*
[28] *Id.* at 7-8.
[29] Alex Lee, *This Ugly T-Shirt Makes You Invisible to Facial Recognition Tech*, Wired, (May 11, 2020, 06:00 AM), https://www.wired.co.uk/article/facial-recognition-t-shirt-block
[30] *Id.*
[31] Draft NISTIR 8269, *supra* note 12, at 8.

conclusions from these observations about its internal workings).[32] In a gray box situation, the adversary will have some limited knowledge of the training model.[33] Under these situations of limited information, there is the potential for significant unintended outcomes due to uncertainty of how the AI systems operate—this could have significant effects on the legal analysis of such attacks in a black- or gray-box environment, as discussed below.

B. What is the legal framework for these attacks on data?

Many scholars are focused on attempting to predict the future of AI and its applications to warfare.[34] In the comparison of AI to electricity—that is, an enabling technology that allows for significant improvements when applied to different contexts—a critical question is how military organizations will apply this enabling technology in their combat operations? Everyone seems to agree that AI will lead to "profound" changes in warfare.[35] However, much will be dependent upon the nature and pace of progress and how different militaries elect to implement this technology. Much of the current scholarly literature focuses on the potential for an automated system that selects and engages targets without human involvement (so-called killer AI or killer robots).[36] While true "Terminator-style" AI is not likely in the foreseeable future, increasing AI support to the targeting process poses the potential for hyper-fast and hyper-accurate targeting that would overwhelm the enemy and result in a decisive and rapid victory.[37] In this context, countermeasures such as data poisoning or data evasion that degrade or slow enemy AI targeting systems will become quite useful.

This chapter identifies and reviews three main legal concerns that should be considered and addressed prior to the deployment of data poisoning attacks on the battlefield. First, how should data poisoning activities that take place prior to armed conflict be regulated? As discussed in the overview of the Draft NISTIR 8269 framework, a critical vulnerability for attack is during the training phase of the machine learning agent. These training phase preparations and corresponding data attacks will likely take place prior to armed conflict—but with impacts that will carry forward into the armed conflict. How does the law regulate such pre-conflict activity? What legal provisions are in place to provide appropriate limits?

Second, attacks on data in armed conflict will likely take place in a highly uncertain, even speculative, environment. While military operators engaged in data poisoning and data evasion will have a specific military objective in mind (specifically, to target the data driving the enemy's algorithms in order to degrade the enemy's AI targeting systems or render them inoperable), the practical effects on the battlefield are difficult if not impossible to predict, particularly in a black box or gray box environment. Artificial intelligence systems hobbled by data poisoning attacks may malfunction in unpredictable ways, which may result in unintended consequences to civilians. How does a data

[32] *Id.*

[33] *Id.*

[34] *See generally*, JAMES JOHNSON, ARTIFICIAL INTELLIGENCE AND THE FUTURE OF WARFARE (2021); AI AT WAR (Sam J. Tangredi & Geoorge Galdorisi, eds., 2021); LOUIS A. DEL MONTE, GENIUS WEAPONS: ARTIFICIAL INTELLIGENCE, AUTONOMOUS WEAPONRY, AND THE FUTURE OF WARFARE (2018).

[35] Michael N. Schmitt & Jeffrey S. Thurnher, *"Out of the Loop": Autonomous Weapons Systems and the Law of Armed Conflict*, 4 HARVARD NAT. SEC. J. 231 (2013).

[36] *See* LOSING HUMANITY, *supra* note 16.

[37] ALEXANDER KOTT, U.S. ARMY RESEARCH LABORATORY, ARL-TN-0901, GROUND WARFARE IN 2050: HOW IT MIGHT LOOK, 9 (2018).

poisoning attacker engage in discrimination, proportionality and precautions analyses in such instances where outcomes cannot be predicted?

Third, how do the perfidy rules apply to data evasion attacks and the associated digital camouflage that armed forces will likely adopt to foil AI systems? With current technology, it is not that difficult to confuse or trick an AI system with minimal "digital camouflage" such as adding small pieces of tape to foil identification of a stop sign.[38] While military AI that is relied upon for targeting is likely to be more robust prior to operational deployment, one would expect adversaries to continue to focus on similar simple techniques to foil enemy targeting. Without knowing the inner workings of the enemy system, there is potential for "inadvertent" use of protective markings—situations where the AI/ML system mis-classifies the enemy as a protected person or object due to evasion techniques.[39]

The pace of technological change has made it increasingly difficult for legal regimes to effectively regulate behavior, and the LOAC is just one body of law that has faced such challenges. Increased incorporation of big data and artificial intelligence is likely to make this problem worse. As will be seen in the discussion of these three issues, there are gaps and lacunae in the law which create real risks for some parties to take advantage of these "gray zones." On the other hand, the *lex lata* does give us a jumping off point for beginning to think about how to regulate data poisoning and evasion attacks and suggest approaches for future development of the law. The following three sections will examine each of these major areas in turn.

III. Data Poisoning Attacks Prior to Armed Conflict: *Jus pre Bello*?

One of the first features that is observed of data poisoning, especially if it takes place during the training phase of machine learning, is that such activities may take place well before the occurrence of hostilities. The LOAC generally applies only after armed conflict has commenced. If this is the case, it is unclear whether the LOAC to governs pre-conflict data poisoning activities.[40] Such a dynamic creates a unique situation in warfare—the prospect of operations against data, conducted wholly before the armed conflict and without physical impact before the armed conflict, yet causing physical effects during the armed conflict. This section will address three questions that are raised by this possibility: (1) does the LOAC apply to data poisoning taking place before armed conflict?; (2) does data poisoning itself constitute the start of armed conflict?; and (3) could data poisoning be regulated by the LOAC if effects manifest during conflict?

Building on the Scenario chapter, let us suppose that Outlandia engages in a successful data poisoning attack on Newtropia's AI Targeting System (NAITS) prior to the conflict. Outlandian officials do so by injecting compromised data into the training data, which in turn undermines the algorithm's effectiveness and degrades the system's overall reliability. The NAITS performs well enough to pass Newtropian reliability tests prior to the fielding of the weapon, but the system's designers notice that NAITS is not as effective as their projections had indicated. Suppose that post-conflict analysis determines that this decrease in the NAITS effectiveness is due to Outlandian data poisoning efforts.

[38] Lee, *supra* note 29.

[39] Recognizing, of course, there is an intent component to perfidy.

[40] Note that other legal regimes may apply to such situations, such as International Human Rights Law, *jus ad bellum*, international law related to sovereignty, or domestic law. These legal regimes will not be addressed in this chapter.

However, this difference is not a mere statistic, as it directly led to an increase in mis-targeting and corresponding increase in civilian casualties.

Assuming an extreme scenario in which Outlandian officials are aware that their pre-conflict data poisoning attack would result in excessive civilian casualties (in relation to the anticipated military advantage of “attacking” the data) and yet continue with the attack. Would their actions violate the LOAC where the data poisoning activities entirely took place prior to the conflict?[41] A similar problem would arise where Outlandian officials knew of feasible precautions that would reduce civilian casualties but failed to adopt the precautions. Because Outlandia would have taken such actions *before* the armed conflict, the LOAC would not seem to apply, and Outlandian officials would not likely be deemed culpable under the LOAC .[42] The remainder of this section will outline the possible approaches to resolve this potential legal gap.

A. Has Armed Conflict Been Triggered?

The well-known trigger for the applicability of the Geneva Conventions and the LOAC generally is armed conflict. Common Article Two of the Geneva Conventions establishes the trigger: “the present convention shall apply to all cases of declared war or of any other armed conflict which may arise between two or more of the High Contracting Parties, even if the state of war is not recognized by one of them.”[43] While the primary focus of this provision was extending the applicability of the LOAC beyond cases of declared war, the commonly understood definition of armed conflict arising immediately after Geneva (e.g., “[a]ny difference arising between two States and leading to the intervention of armed forces”[44]) does not shed much light on the level of armed conflict required to trigger this threshold.

Resolving the ambiguity surrounding this issue is the focus of the 2010 International Law Association Final Report on the Meaning of Armed Conflict in International Law.[45] A data poisoning

[41] Recalling that Common Article 2 of the Geneva Conventions designated the timeframe to which the Conventions apply: “In addition to the provisions which shall be implemented in peacetime, [none of which are apparently applicable to the present scenario] the present Convention shall apply to all cases of declared war or of any other armed conflict which may arise between two or more of the High Contracting Parties…”

[42] Again, highlighting the limit of this chapter to the Law of Armed Conflict. Another argument that Outlandia might make is that the superseding and intervening act of Newtropia to deploy the NAITS despite its degraded effectiveness negated Outlandia’s legal culpability. Outlandia would likely claim that they relied on Newtropia to act in good faith and comply with IHL by taking the system offline if its effectiveness were degraded to the point where its deployment would lead to excessive civilian casualties. This argument is considered in the proportionality analysis *infra*.

[43] Common Article 2 of the Geneva Conventions (1949). Note that this portion is preceded by a qualifier, “In addition to the provisions which shall be implemented in peacetime…” *Id.* However, these are limited in scope to such things as marking protected and cultural property and training forces on the law of war. There appear to be no provisions applicable to peacetime that are relevant to the data poisoning problem.

[44] INT’L COMMITTEE OF THE RED CROSS, COMMENTARY I GENEVA CONVENTION FOR THE AMELIORATION OF THE CONDITION OF THE WOUNDED AND SICK IN ARMED FORCES I N THE FIELD 32 (Jean S. Pictet ed., 1952).

[45] INT’L L ASS’N, FINAL REPORT ON THE MEANING OF ARMED CONFLICT IN INTERNATIONAL LAW (2010) (hereafter ILA Armed Conflict Final Report). Note that while a major concern in the report was determining when Non-International Armed Conflicts met the armed conflict threshold, the committee focused on a general definition of armed conflict and considered both Common Article 2 International Armed Conflict as well as Common Article 3 Non-International Armed Conflict. *Id.* at 3 n.7.

scenario is not likely to meet the criterion of intensity ("hostilities must reach a certain level of intensity to qualify as an armed conflict"[46]). Among the factors that the ILA cites to assess intensity based on their review of custom, commentary and judicial opinions, none seem to even reach the level of a colorable claim for intensity in a data poisoning situation like that outlined in the Scenario chapter. Specifically, the ILA suggests the following factors to assess intensity: 1. "number of fighters involved"– in the case of data poisoning, none would actually be exchanging kinetic attacks, although there may be several involved in the data poisoning operation; 2. "type and quantity of weapons used"—in this case there would be cyber operations but with no immediate physical effects; 3. "the duration and territorial extent of fighting"—no fighting would be occurring because data poisoning is surreptitious, all operations associated with the data poisoning would take place in cyberspace with no physical effects prior to the commencement of armed conflict; 4. "extent of destruction of property"—no physical property would be destroyed, albeit data might be altered and/or deleted; 5. "displacement of the population"—none in this case; 6. "involvement of the Security Council or other actors to broker cease-fire efforts"—again, not likely.[47] In a situation like that described in the Scenario chapter, it is hard to envision a possible basis to make a good faith claim that hostilities have commenced such that the Common Article 2 Armed Conflict threshold has been triggered.

B. Does data poisoning standing alone meet criteria for armed conflict?

Another option might be to consider whether data poisoning activities, standing alone, serve as the marker of the commencement of armed conflict. Because the current position of the Tallinn 2.0 experts is that an attack against data, standing alone, does not qualify as a cyber attack,[48] such a conclusion is unlikely. With data classified as intangible, and not a physical object whose destruction/deletion would implicate the *jus ad bellum-jus in bello* legal frameworks, it is hard to imagine a situation where the surreptitious insertion of fake data or manipulation of existing data would be sufficient standing alone to trigger an armed conflict. Even under the more expansive definition of attack considered by the International Group of Experts—specifically, cyber operations "result[ing] in large-scale adverse consequences"— data poisoning would not qualify. It is unlikely that data poisoning would have such large-scale adverse effects prior to the commencement of conflict.[49] That said, this position has proved controversial in the *jus in bello* context, with calls to expand the definition of cyber attacks to include attacks on data.[50] At this point, it does not appear that that custom has developed to the point where a conclusion could be drawn that an attack on data would either qualify as a use of force (in the *jus ad bellum* context) or cyber attack (in the *jus in bello* context). The prospect of data poisoning, however, serves as another layer of challenges in conceptualizing and articulating what cyber activities constitute a use of force or armed attack.

C. Continuing Crimes Doctrine as Vehicle for Legal Regulation

[46] *Id.* at 29.

[47] These factors are listed in the ILA Armed Conflict Final Report. *Id.* at 30.

[48] TALLINN 2.0, *supra* note 9, at 416.

[49] *See id.* at 418 (considering and rejecting the position that cyber operations resulting in "large-scale adverse consequences" qualify as an attack under the law of armed conflict).

[50] Kubo Mačák, *Military Objectives 2.0: The Case for Interpreting Computer Data as Objects under International Humanitarian Law*, 48 Israel Law Review 55 (2015).

Another possible vehicle for regulating pre-conflict data poisoning would be the doctrine of continuing crimes. This doctrine relates to "a breach of a prohibition over a period of time."[51] An example of such a crime in International Criminal Law is enforced disappearance, in which the crime continues so long as the whereabouts of the disappeared person are not released.[52] A similar rule exists for continuing State acts under the Draft Articles of State Responsibility.[53] In the common law within the United State, the doctrine of continuing effects also generally operates to extend or toll statutes of limitation that have expired for tort or criminal liability.[54] If a sufficient doctrine of continuing crimes could be identified, it might serve to encompass data poisoning taking place prior to armed conflict and render such activities subject to the LOAC.

Significant challenges would exist, however, to extending this principle to allow for a LOAC violation prior to armed conflict. While the Draft Articles of State Responsibility allow for continuing violations, they also explicitly disallow the retroactive application of the law: "An act of a State does not constitute a breach of an international obligation unless the State is bound by the obligation at the time the act occurs."[55] Similarly, the principle of *nullum crimen sine lege, nullum poene sine lege* would operate to restrict application of retroactive criminal liability for war crimes. A practical example of these principles is the Rome Statute, which applies only to offenses "committed after the entry into force of [the] Statute."[56] Even continuing crimes commencing prior to the entry of force of the Rome Statute are not clearly addressed by the Statute and would be subject to the limitations of *nullem crimen sine lege*.[57]

To better parse whether a continuing crimes/effects theory is viable, it will be necessary to examine the practical aspects of the specific type of data poisoning attack being used. The specific details of the poisoning attack will help to answer the question whether the poisoning is an ongoing event that might qualify as "continuing" or whether it is instead a discrete event with a defined end point. A training phase data poisoning attack would likely qualify as the former where the poisoning concludes at the end of the training phase prior to the AI system becoming operational. Another consideration that might serve to classify a data poisoning attack as a "continuing" act is the degree it is possible for a data poisoning attacker to reverse the effects of data poisoning through removing the poisoned data. Generally, one might expect that the data poisoning would set off an irreversible sequence of events that cannot be remedied by an *ex post* fixing of the data (if fixing the data were even possible). In assessing whether an attack would be continuing in nature, it will be necessary to examine the details of the specific type of attack, but in many cases it appears doubtful that an argument can be made for the poisoning to qualify as a continuing act.

[51] Alan Nissel, *Continuing Crimes in the Rome Statute*, 25 MICH. J. INT'L L. 653, 661-62 (2004).

[52] *Id.* at 654.

[53] International Law Commission, Draft Articles on Responsibility of States for Internationally Wrongful Acts, arts. 14 & 30, November 2001, Supplement No. 10 (A/56/10), chp.IV.E.1 [hereinafter Draft Articles of State Responsibility].

[54] *See generally,* Kyle Graham, *The Continuing Violations Doctrine*, 43 GONZAGA L. REV. 271 (2007).

[55] Draft Articles of State Responsibility, *supra* note 53, art. 13.

[56] Rome Statute of the International Criminal Court, *opened for signature* July 17, 1998, art. 11(1), 37 I.L.M. 999, 1010 (entered into force Jul. 1, 2002) [hereinafter Rome Statute].

[57] Nissel, *supra* note 51, at 656, 687.

Another significant objection that must be resolved before travelling further down the continuing crimes path is the fact that there is no LOAC violation that would apply to invalidate the original, pre-conflict data poisoning because the LOAC is not yet in effect. The doctrines noted above pre-suppose that the misconduct in question violated then-existing law. In other words, the act of data poisoning, which might be legal before armed conflict, might be transformed *ex post facto* into a LOAC violation and perhaps a war crime, by virtue of the superseding and intervening cause of armed conflict commencing and the corresponding applicability of the LOAC.[58] This objection poses significant challenges to retroactive application of the LOAC, making the applicability of the continuing crimes doctrine unlikely.

D. Applicability of Martens Clause

The previous discussion leaves the Martens Clause as the most viable option for providing some meaningful legal regulation to pre-conflict data poisoning. While the Clause exists in different forms depending on the treaty in which it is found, the original iteration provides:

> "Until a more complete code of the laws of war has been issued, the High Contracting parties deem it expedient to declare that, in cases not including in the Regulations adopted by them, the inhabitants and the belligerents remain under the protection and rule of the principles of the law of nations, as they result from the usages established among civilized peoples, from the laws of humanity, and the dictates of the public conscience."[59]

Interpreting this clause has presented a challenge as there is no settled agreement on interpreting the reach and applicability of the Martens Clause. This chapter will not attempt to examine the various interpretive approaches in detail or defend a detailed theory of legal regulation of pre-conflict data poisoning attacks via the Martens Clause. Instead, this section will explore whether the Martens Clause might provide a method of regulating pre-conflict data poisoning attacks.

While scholars have asserted different interpretive approaches to the Martens Clause,[60] only one interpretation provides any real inroads towards regulating pre-conflict data poisoning. This theory

[58] Note that this analysis is restricted to the Law of Armed Conflict. The applicability of other bodies of law, such as Human Rights Law, are not under consideration in this chapter.

[59] 1899 Hague Convention (II) Respecting the Laws and Customs of War on Land with Annex of Regulations, preamble, July 29, 1899, 32 Stat. 1803, 1 Bevans 247.

[60] Four interpretive approaches can be ruled out at the outset as not providing substantial guidance to the problem of pre-conflict data poisoning. (1) The *a contrario* argument states that the Martens Clause merely highlights that existing customary law continues to regulate armed conflict in the absence of an authoritative treaty provision. Antonio Cassesse, *The Martens Clause: Half a Loaf or Simply Pie in the Sky*?, 11 EUR. J. OF INT'L L. 187, 192 (2000). This approach is not helpful due to the absence of customary Law of Armed Conflict provisions that might serve to regulate data poisoning pre-conflict. (2) Another tack is to argue that the Clause serves to elevate the value of *opinio juris* and loosen the requirements for concordant state practice in the field of International Humanitarian Law in order to establish custom, *Id.* at 214; Theodor Meron, *The Martens Clause, Principles of Humanity and Dictates of Public Conscience*, 94 AM. J. OF INT'L L. 78, 88 (2000). This view does not assist in the absence of *opinio juris* and state practice in the arena of data poisoning in armed conflict. (3) Another interpretation is that the Martens Clause is interpretive gloss, "merely" serving to elevate considerations of humanity in close situations. Cassesse at 212; Meron at 88. This interpretive gloss would be of little effect if the trigger of armed conflict has not been met, as discussed earlier. (4) Finally, one might be able to import the argument forwarded by Human Rights

asserts that the Martens Clause recognizes (or creates) a body of law with an expansive applicability, which might serve as a possible means of regulating pre-conflict data poisoning. The strong version of this theory—concluding that the Clause raises principles of humanity and public conscience to the level of general principles of international law[61]—is likely not to be accepted as a basis for regulating data poisoning, as there does not appear to be widespread acceptance of this position. However, a weaker formulation of this argument is more defensible. Schmitt and Thurnher offer the best expression of this formulation: "By its own terms, though, the clause applies only in the absence of treaty law. In other words, it is a failsafe mechanism meant to address lacunae in the law; it does not act as an overarching principle that must be considered in every case."[62] In fact, this statement seems to directly address the problem created by pre-conflict data poisoning—it is a situation not governed by existing treaty and appears to be a lacuna in the law.

Under this approach, inquiry must be made to determine what law would apply to pre-conflict activities and the source of this law. There would be no customary law, as there is no state practice and *opinio juris* addressing pre-conflict data poisoning. However, a case could be made that principles of humanity woven into the LOAC could provide the outlines for a legal framework. Three threads or principles from the LOAC, woven together, outline the potential case that, under the Martens Clause, the LOAC should apply to pre-conflict data poisoning. First, Common Article 3 contains very expansive language: "[Noncombatants] shall *in all circumstances* be treated humanely…" and "… the following acts are and shall remain prohibited *at any time and in any place whatsoever*…". While the opening sentence of Common Article 3 contains the "armed conflict not of an international character" qualifier, the expansive language cited in the previous sentence ("in all circumstances" and "at any time and in any place whatsoever") suggests that there might be some aspects of Common Article 3 that extend beyond armed conflict. This position is further buttressed by the ICJ *Corfu Channel* decision, in which the ICJ stated that the violations of Albania were "obligations based, not on the Hague Convention of 1907, No. VIII, which is applicable in time of war, but on certain general and well-recognized principles, namely: elementary considerations of humanity, even more exacting in peace than in war…"[63] The United Nations Security Council used a second, parallel approach to condemn the targeting of civil aircraft as a violation of the LOAC even though the targeting took place in flight outside of the armed conflict environment, where such actions were described as "being incompatible with elementary considerations of humanity."[64] The third thread buttressing such a position was the ICJ's statement in the *Advisory Opinion on the Legality of the Threat or Use of Nuclear Weapons.* The Court stated there that that the Martens Clause "proved to be an effective means of addressing the rapid evolution of military technology."[65]

Watch regarding the use of autonomous weapons; specifically, that certain weapons systems should be prohibited under the Martens Clause *ab initio* if they are determined to be contrary to humanity or public conscience. HUMAN RIGHTS WATCH, MAKING THE CASE: THE DANGERS OF KILLER ROBOTS AND THE NEED FOR A PREEMPTIVE BAN, 14-17 (2016), *available at* https://www.hrw.org/sites/default/files/report_pdf/arms1216_web.pdf. However, it would need to be shown that the use of data poisoning is contrary to the principles of humanity or public conscience.

[61] Meron, *supra* note 60, at 80-82.

[62] Schmitt & Thurnher, *supra* note 35, at 275.

[63] Corfu Channel (UK v. Alb.) (Merits), 1949 I.C.J. Rep. 4, 22 (Apr. 9).

[64] S.C. Res. 1067, ¶ 6 (July 28, 1996).

[65] Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, ¶ 78 (July 8).

These three threads—the expansive language of Common Article 3, the recognition of considerations of humanity as existing in both peacetime and war, and the ICJ's identification of the Martens Clause as a tool to help address technological advances—provide an outline of a theory to regulate pre-armed conflict data poisoning. This outline is by no means unassailable, but it provides a useful starting point. In order to make a stronger case, one would need to better address the counter-argument that the Martens Clause is merely a gap-filler within armed conflict (with no applicability outside of an armed conflict situation). One way to address this argument is to examine the degree to which the LOAC serves as *lex specialis*, with the Martens Clause referring to (and perhaps recognizing to a certain degree) the *lex generalis*, consisting of the principles of humanity universally applicable as noted in the *Corfu Channel* opinion. Given the increased blurring of the lines between conflict and non-conflict situations, further scholarly work in addressing how the Martens Clause might address these gray zones between peace and war would be helpful.

The next question to consider is the substantive content of the legal framework that the Martens Clause would import into a pre-conflict scenario. This topic itself could be the subject of an extensive article and goes beyond the scope of this chapter. In addition, as will be discussed in the next section, even the legal framework governing use of data poisoning in armed conflict contains gaps, primarily due to the indirect nature of the attack and the fact that the adversary has the ability to prevent bad outcomes by taking the compromised AI system offline. One possible guide, however, might be the UN Special Rapporteur's report on human rights in occupied Kuwait, which stated that the "elementary considerations of humanity" referenced in the Martens Clause incorporated these three principles: "(i) that the right of parties to choose the means and methods of warfare, i.e. the right of the parties to a conflict to adopt means of injuring the enemy, is not unlimited; (ii) that a distinction must be made between persons participating in military operations and those belong to the civilian population to the effect that the latter be spared as much as possible; and (iii) it is prohibited to launch attacks against the civilian population as such."[66] This list of elementary considerations of humanity is a good start, but further research is needed on whether and how the Martens Clause might apply to pre-conflict data poisoning.

IV. Data Poisoning During Armed Conflict: Targeting Principles in an Uncertain Environment

Once conflict has commenced, the traditional targeting legal analysis becomes more central. It is important at the outset to highlight that the attack on big data outlined in the Scenario chapter operates as a "hack"—a back door means to disrupt or degrade the artificial intelligence. This fact creates vexing challenges for prospective attackers and their legal advisors due to the uncertain effects and attenuated nature of the data poisoning attack.[67] The data poisoning attacker definitely seeks an effect—to degrade or deny the enemy use of their AI system as it targets their forces. But *how* data poisoning achieves this effect is murky at best and unknowable at worst. In some ways, data poisoning is a shot in the dark. This shot in the dark may be acceptable (and not considered an indiscriminate attack) if the data poisoning attack consisted of a "mere" cyber operation targeting data with no

[66] Report on the Situation of Human Rights in Kuwait under Iraqi Occupation, para. 36, U.N. Doc. E/CH.4/1992/26.

[67] For purposes of this section, I will operate from a working assumption that a data poisoning attack constitutes a cyber attack and is therefore governed by the Law of Armed Conflict.

immediate kinetic effects.[68] On the other hand, the analysis may be different if there is the potential for the data poisoning attack to result in inaccuracies with the AI system, resulting in the targeting of civilians and civilian objects. This section examines targeting principles in the highly uncertain data poisoning environment (from the starting position that a data poisoning attack in armed conflict constitutes a cyber attack and thus governed by LOAC targeting rules), by examining in turn whether data poisoning is an indiscriminate method of warfare, how to conduct a proportionality analysis, and the applicability of precautions in the attack.

A. Is a data poisoning attack an indiscriminate means or method of warfare?

To provide context in analyzing whether data poisoning might constitute an indiscriminate attack, an analogy might be made to a GPS-guided precision bomb. If the targeted military elected to engage in GPS-jamming of the precision bomb, thus rendering the munition a "dumb" bomb with the potential for landing in civilian areas, is the GPS-jamming operation an indiscriminate attack?[69] Similarly, does a data poisoning attacker who disrupts or degrades an AI targeting system cause an indiscriminate attack?

From a strictly textual analysis of the three types of indiscriminate attacks found in Article 51(4) of Additional Protocol I,[70] an argument can be made that data poisoning does not constitute an indiscriminate attack. Because the data poisoning attacker is directing the poisoned data to a specific military objective,[71] the enemy's AI systems, it would not qualify as the first type of indiscriminate attack ("those which are not directed at a specific military objective"). The ultimate objective of the data poisoner is likely to reduce the number and effectiveness of AI-supported attacks on its own troops, or, in an extreme case, force the enemy to shut down its AI-targeting systems altogether. Nor would data poisoning likely constitute the second type of indiscriminate attacks, "those which employ a method or means of combat which cannot be directed at a specific military objective." The attacker can feed poisoned data by directing it at a specific military objective with little danger of spillover.[72] Unlike a worm or virus, which carries the risk of propagation beyond the intended military network being attacked, data poisoning is likely to be tailored to the system being poisoned, does not propagate, and does not have likely carry-over affects to other AI systems utilizing different use-cases.

[68] *See generally*, Michael N. Schmitt, *Rewired Warfare: Rethinking the Law of Cyber Attack*, 96 Int'l Rev. Red Cross 189 (2014) (outlining the debate over the two approaches to defining what cyber operations constitutes an attack, whether an attack on data constitutes an attack and describing the Tallinn Manual deliberations concluding that an attack on data, standing alone and without physical effect or effect on functionality would not constitute a cyber attack); *see also* TALLINN MANUAL 2.0, *supra* note 9, at 416 (stating the same conclusion).

[69] Brigadier General (Retired) David Wallace, Professor Emeritus, U.S. Military Academy, provided this analogy.

[70] Protocol Additional to the Geneva Conventions of 12 August 1949, art. 51 (June 8, 1977) [*hereinafter,* Additional Protocol I]. The three types or attacks are: 1. "those which are not directed at a specific military objective;" 2. "those which employ a method or means of combat which cannot be directed at a specific military objective;" or 3. "those which employ a method or means of combat the effects of which cannot be limited as required by this Protocol." *Id.* at art. 51(4).

[71] *See id.* at art. 51(4)(a) (defining indiscriminate attacks as "those which are not directed at a specific military objective.").

[72] *See id.*at art.51(4)(b) (defining indiscriminate attacks as "a method or means of combat which cannot be directed a specific military objective.")

The third type of indiscriminate attacks, "those which employ a method or means of combat the effects of which cannot be limited as required by this Protocol,"[73] does create potential issues. The expected outcome of a data poisoning attack would be to degrade performance in the AI system—most likely in the form of increased processing times (causing delayed targeting decisions) and lower accuracy (causing an increased the likelihood of errors, to include the mistaken targeting of noncombatants). The most extreme outcome would be a situation in which the AI system targets wholly at random, creating the potential for truly indiscriminate and potentially devastating effects. Any such outcome is a second-order effect and not a direct effect of the data poisoning—the subsequent malfunction of the enemy AI system is what causes the adverse effects. In addition, this effect can be limited if the adversary elected to take the AI system offline either due to the observation of the malfunction by the AI system operators or due to notification of the data poisoning attack's effects by the data poisoning attacker. Given the attenuation of effects and potential to limit effects, the better answer is to conclude that a data poisoning attack is not indiscriminate.

Instead, the concerns for civilian collateral damage are better suited for regulation by the principle of proportionality. In such a proportionality analysis, the data poisoning attacker must evaluate the expected civilian losses against the anticipated military advantage of the GPS-jamming hypothetical presented earlier or data poisoning. The proportionality route offers several advantages, as (1) it does not require the combatants to remain so-called "sitting-duck" targets and (2) it also takes into account the individual facts and circumstances of the GPS jamming or data poisoning operation (e.g., GPS-jamming taking place in the desert vs. an urban environment). Further, the fact that potential adverse effects of data poisoning can be controlled by the AI-system operator through turning the system off can effectively be weighed in a proportionality analysis, lending further weight in favor of proportionality approach as the better way to evaluate the legality of data poisoning. The next section will address the nuances of that particular issue.

B. How to Apply Proportionality to the Data Poisoning Context?

Moving now into the nuts and bolts of a proportionality analysis, some of the features of data poisoning serve as significant challenges to the commander attempting to engage in a proportionality assessment prior to attack. Under the traditional Additional Protocol I formulation, which prohibits an attack "which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated,"[74] both sides of this analysis pose challenges in the data poisoning context. What is "the concrete and direct military advantage anticipated"?[75] A data poisoning attacker would hope that the attack renders the enemy AI inoperable, although more realistically expects a degradation in the system's effectiveness. The outcome cannot be known at the outset, however. What civilian collateral effects "may be expected"?[76] This second question is even more difficult, as it is more attenuated from the immediate purpose of the data poisoning attack. Finally, as in all combat, "the enemy gets a vote." Here, how do the enemy's actions in operating the system work to mitigate

[73] *Id.* at art. 51(4)(c).

[74] *Id.* at Art. 51(5)(b).

[75] *Id.*

[76] *Id.*

potential collateral civilian damage? This section will examine first the traditional proportionality calculus and then will turn to how enemy activity might affect proportionality.

1. Proportionality Analysis Challenges. Data poisoning, as a result of its indirect, uncertain and attenuated nature, will challenge the commander's ability to conduct a proportionality analysis. In the first place, the data poisoning attacker cannot be sure of the ultimate effects that will be accomplished through the attack. An AI targeting system being attacked via data poisoning would be a military target. The desired effect of a data poisoning attack is to degrade the effectiveness of the targeting system—generally in terms of accuracy or timeliness. A degraded system will likely operate at a slower speed, thus giving the data poisoning attacker an advantage, perhaps even a decisive advantage. For example, a reduction in accuracy will likely provide a great advantage if the data poisoning reduces the enemy AI system to 72% accuracy instead of 98% accuracy. The data poisoning attacker's objective is to achieve a significant effect such as degrading the system to the point of ineffectiveness, but less significant effects such as minor degradation, or an effect lasting a short period of time, or no effect at all, are distinct possibilities. Due to the uncertainty of effect referenced in the previous paragraph, the concrete and direct military advantage is difficult to assess with any level of certainty.

Of course, by degrading AI targeting, data poisoning is likely to come with a cost on the collateral effects side. As an AI targeting system loses accuracy, one must consider the real-world targeting effects: whom or what is being targeted in lieu of military targets? If combat takes place in a remote area, then such concerns are minimal. However, combat in populated areas become more problematic. The black box nature of most AI systems will further complicate matters, further limiting the ability to predict potential adverse effects. Much like the anticipate military advantage, the expected civilian collateral effects will be difficult to assess in advance.

The debate over the interpretation of "may be expected" and the role of reverberating effects further hinders satisfactory resolution of these matters.[77] The nature of data poisoning is such that the most likely outcome would be a chain of events that results in collateral damage (e.g., poisoned data during the training phase alters the machine learning algorithm, which in turn causes errors in targeting). This chain of causation potentially extends beyond what one may expect a commander to anticipate, depending on one's interpretation of how causation applies in the proportionality context. On the other hand, while there is debate over interpretations of causation and foreseeability, a reasonable person or reasonable commander would likely expect some degree of civilian collateral effects to result from a successful data poisoning attack. Because of this fact, a proportionality analysis should be conducted by the data poisoning attacker.[78]

[77] *See* Ian Henderson & Kate Reece, *Proportionality under International Humanitarian Law: The Reasonable Military Commander and Reverberating Effects*, 51 VAND. J. TRANSNAT'L L. 835, 846-54 (2018) (outlining the arguments and considerations regarding foreseeability and direct vs. indirect effects when assessing proportionality); Isabel Robinson & Ellen Nohle, *Proportionality and Precautions in Attack: The Reverberating Effects of Using Explosive Weapons in Populated Areas*, 98(1) INT'L REV. RED CROSS, 107, 116 (2016) (concluding that "certain reverberating effects" must be considered when assessing proportionality); YORAM DINSTEIN, THE CONDUCT OF HOSTILITIES UNDER THE LAW OF INTERNATIONAL ARMED CONFLICT 159 (3rd ed., 2016) (restricting analysis to direct effects); Eric Talbot Jensen, *Cyber Attacks: Proportionality and Precautions in the Attack*, 89 INT'L L. STUD. 198, 208 (2013) (indirect effects should not be factored unless "expected").

[78] *See, e.g.,* TALLINN MANUAL 2.0, *supra* note 9, at 472 (indicating that indirect effects should be considered as collateral damage where the damage is "expected" by those engaging in a cyber attack).

A possible solution to the problem of not being unable to precisely predict the outcome is to engage in a sliding scale proportionality analysis, analyzing the range of possible outcomes. For example, relatively minor effects of data poisoning on the enemy AI is likely to have relatively small civilian collateral damage. Similarly, a significant military advantage such as rendering the enemy AI inoperable would have great military advantage but would also carry with it a much higher chance of significant collateral damage. One difficulty to this approach would be if some outcomes were deemed disproportionate and others were not—which outcome would control whether the ultimate proportionality determination? An intuitive possible answer might be to select the outcome which was considered most likely by the data poisoning experts, but there are other approaches which should be considered and addressed. Further research and analysis is needed in applying the proportionality standards to data poisoning.

2. Proportionality and Enemy Responses to Data Poisoning. The other factor complicating the proportionality analysis is the fact that the targeted enemy system itself will be the source of the collateral damage. On the one hand, from a pure causation view, this fact presents the possibility of a superseding and intervening cause the removes responsibility from the data poisoning attacker, because the enemy is the entity that will be conducting the targeting operations. On the other hand, it does not seem advisable to conclude that the attacker lacks all responsibility for this reason. A middle ground seems more appropriate; in such an approach, a proportionality analysis could consider three factors.

The first factor is the degree to which the AI targeting system has a human operator in the loop. This factor intersects the current debate over autonomous weapons systems and the degree of active and continuing human oversight of AI-assisted or AI-directed targeting.[79] If there is a high level of human oversight, the potential for civilian collateral damage will be limited if an AI system malfunctions due to data poisoning. The higher degree of human oversight is likely to mitigate negative outcomes.[80] As a result, a data poisoning attacker could confidently assess a lower likelihood of collateral damage when conducting a proportionality analysis in a human "in the loop" scenario. Alternatively, in a human "out of the loop" situation, with the AI able to make substantive targeting decisions on its own, there is the potential for significant adverse consequences should the AI begin targeting at random as a result of the data poisoning attack. This is due to the significant time lag that is likely to occur before such a situation is identified and remedied. In such instances, the data poisoning attacker should expect a higher level of collateral damage.

Given the controversy surrounding autonomous weapons,[81] this analysis does create a potential problem whereby the entity operating the AI targeting system might be disincentivized to keep a human in the loop. An attack on an unmanned system may be foreclosed because the data poisoning attacker deems to the proportionality risk to be too great. There is a very real possibility of creating a moral

[79] *See generally,* Alan L. Schuller, *At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law*, 8 Harv. Nat'l Sec. J. 379, (2017); Lieutenant Colonel Christopher M. Ford, *Autonomous Weapons and International Law*, 69 S.C. L. Rev. 413, (2017); Marco Sassoli, *Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified*, 90 Int'l L. Stud. 308 (2014); Schmitt and Thurnher, *supra* note 35; Hin-Yan Liu, *Categorization and Legality of Autonomous and Remote Weapons Systems*, 94 Int'l Rev. Red Cross 627 (2012).
[80] *See* Schmitt and Thurnher, *supra* note 35, at 234-43 (outlining the different degrees of human oversight of autonomous systems).
[81] *See generally*, *id.*; Losing Humanity, *supra* note 16.

hazard, protecting those who deploy AI weapons systems without meaningful human oversight, if proportionality is not implemented thoughtfully here. As the law surrounding the use of autonomous AI weapons continues to develop, it may be advisable to limit the reach of this factor in order to further the normative goal of minimizing unmanned weapons systems.

The second factor for consideration is the degree to which human oversight would be able to see the effects as they are occurring in real time. The data poisoner should also consider the degree of control and oversight that the enemy is able to exercise over the AI system. If the AI system operator has good visibility over the system and is readily able to take the AI system offline if the system begins targeting civilians due to the data poisoning, then proportionality concerns will be minimized. On the other hand, if the enemy has limited ability to observe the effects of the AI targeting, for example due to other cyber operations against the AI system, this factor should be considered in the proportionality analysis that the data poisoner conducts.

Finally, the potential collateral effects of a data poisoning attack differ from those of other attacks because the side being poisoned has control over the targeting system and can turn it off should it the AI targeting system become too error-prone (assuming they are able to see the effects of the data poisoning attack). An obvious call would be if a data poisoning attack resulted in an AI targeting system, as a result of the attack, begins to target indiscriminately or at random. A human operator overseeing the system would immediately have the responsibility to take such a damaged system offline. But what about the attack mentioned above where the overall accuracy is reduced from 95% to 72%. Or 62%? Or 47%? At what point is the operator of the AI targeting system obliged to take the system offline? And if the operator continues to rely on a degraded AI system (because presumably they assess that the collateral damage is not excessive in relation to the military advantages of continuing to operate the system), can one maintain that the data poisoner is responsible for the adverse effects against civilian objects? The question that must be considered is whether a data poisoning attacker should be able to rely on the reasonable good faith of their adversary to take the AI system offline if the data poisoning is so effective that it results in either excessive civilian casualties or indiscriminate attacks.

When considering proportionality, the attenuated nature of the downstream effects of a data poisoning attack and enemy control of the AI creates challenges in applying the proportionality legal framework. If we reach a point where data poisoning attacks are actually used in combat, we will have real-world experience from which subsequent data-poisoning attackers will be able to draw in order to conduct a proportionality analysis. Until we reach that point, the unique nature of data poisoning will present challenges to a commander conducting a proportionality assessment. The factors identified in this section should assist the commander in making this determination in a way that mitigates the most serious potential harms.

C. What Precautions in the Attack are feasible for data poisoning?

In the arena of precautions in the attack, which requires that "constant care shall be taken to spare the civilian population, civilians and civilian objects,"[82] data poisoning poses similar challenges. These challenges center around the same uncertainty and attenuation issues discussed in above. However, there are some unique considerations that may be applicable. To that end, I propose a number of guidelines that data poisoning attackers should consider when planning and executing a data

[82] Additional Protocol I, *supra* note 70, art. 57(1).

poisoning attack. I intend these suggestions as a starting point, as the proposals are likely to be of limited utility with the current state of technology. At a minimum, the legal advisor should discuss the principle of precautions in the attack with the cyber operators developing a poisoning attack so precautions could be identified and built into the program where it was feasible.

Suggested Precautions Guideline 1: To the extent feasible, design data poisoning attack with a view towards minimizing civilian casualties. This guideline may be hard to implement in reality at present, but those who conduct future data poisoning attacks may be able to implement this precaution. As discussed earlier, the data poisoning attack is likely to have uncertain, unpredictable and highly variable downstream responses. In addition, the indirect, second- and third- order outcomes are the ones which will potentially result in civilian casualties. That said, in the technical process of developing the data poisoning attack, there may be opportunities to build in features or options or design the software in a way to minimize the possibility of an automated system targeting civilians or civilian objects. While a legal advisor might not be able to meaningfully understand the coding process and other technical aspects of designing such an attack, he or she can outline the requirements for precautions in the attack and engage in a conversation about potential precautions where such efforts might be feasible.

Suggested Precautions Guideline 2: Consider the feasibility of designing a reversal capacity for a data poisoning attack in the event the attack causes the targeted system to become uncontrollable and begin causing civilian attacks. Much like the first guideline, the capacity to reverse the effects of a data poisoning attack may be difficult to achieve with present technology. This would be the case particularly if the poisoned data were introduced into the dataset on a one-time basis (e.g., fire and forget data poisoning). If the data poisoning attacker were able to achieve an ongoing introduction of poisoned data into the enemy's system, then the data poisoning attacker might be under an obligation to attempt to observe the ongoing effects, if such observation were feasible, and turn off the flow of poisoned information if it became apparent that the poisoned data was resulting in civilian attacks that were excessive in relation to the military advantage. This is similar to the suggestion that a commander should engage in continuous monitoring of enemy networks, if feasible, in order to comply with his or her precautions obligations.[83]

Suggested Precautions Guideline 3: Adopt procedures for notifying the enemy if there are indicators that the targeted system is engaging or is about to engage in significant attacks on civilians. Much like the first two guidelines, this potential precaution will likely be of limited utility. First, the operators of the targeted AI system will likely be in a better position to receive indicators of malfunction of this nature. Second, data poisoning or data evasion attacks generally do not involve ongoing monitoring of the enemy systems, although perhaps the data poisoning attacker may be in a position to see physical effects and link them to ongoing data poisoning operations—thus providing an opportunity for the data poisoning attacker to see effects being caused by the data poisoning. This approach would not be without controversy, however. Notifying an enemy that compromised data was adversely affecting their AI systems would provide the enemy an opportunity to scrub the data and unleash the re-tooled AI system on the armed forces of the data poisoner. This scenario would require close

[83] Jensen, *supra* note 77, at 202-203.

consultations between the data engineers and legal advisors to decide that such a notification would be feasible.

Despite the limitations discussed above, these suggested guidelines are a good starting point for building precautions into prospective data poisoning attacks. At a minimum, they can facilitate a discussion between the data engineers designing prospective attacks and the operational lawyers advising the commander. These examples may inspire other possible ideas for minimizing civilian casualties that would be both feasible and also have a practical ability to limit civilian casualties.

V. Data Evasion During Armed Conflict: Perfidy Considerations

Once active hostilities begin in an armed conflict scenario, the focus of the efforts to undermine enemy AI will likely occur through data evasion, which seeks to poison the data being evaluated by the operational AI system. In such a battlefield, one could expect to see video feeds, motion detection sensors and radio intercepts being piped to the big data servers, like streams feeding a reservoir. These streams have the potential to modify the machine learning, as the system continuously "learns" and updates itself based on the newly received data. Just as an individual can wear a shirt with digital patterns that causes the AI to mis-classify the individual,[84] we should expect soldiers to seek to fool enemy AI targeting systems with similar random symbols (or perturbations, to use the technical terminology). Like the other types of data poisoning, use of this technique creates new legal issues, particularly centered around perfidy. The armed forces seeking to fool the enemy AI will not necessarily know *how* the AI is mis-classifying the otherwise-legitimate target. They would just know that, once they poison the data that powers the enemy AI, the AI is no longer targeting their forces (and that would likely be enough from their point of view). The enemy might then seek to take advantage of the patterns used by the poisoned AI system, raising questions about perfidy. For example, the patterns adopted by the parties might cause the AI to conclude that the soldiers are protected parties, such as medical personnel or members of the Red Cross. The rules of perfidy, which one prominent scholar has critiqued as moving from a "general principle understood into a technically-bound, law of war prohibition,"[85] are not clearly designed to apply to effectively regulate these situations.[86]

As an example, suppose that all International Committee of the Red Cross (ICRC) personnel carry a distinctive messenger bag while engaged in their field work during an armed conflict, and the machine learning of one party's targeting system relies on the presence of the messenger bag to conclude that a person is a member of the Red Cross and therefore not targetable.[87] If the adversary is able to learn of

[84] Lee, *supra* note 29.

[85] Sean Watts, *Law-of-War Perfidy*, 219 MIL. L. REV. 106, 167 (2014).

[86] Generally speaking, perfidy is defined as "any attempt to gain the enemy's confidence by assuring his protection under the law of war, while intending to kill, wound, or capture him." GARY D. SOLIS, THE LAW OF ARMED CONFLICT 458 (2d. ed., 2016). A ruse is considered lawful and is defined as "a deceit employed in the interest of military operations for the purpose of misleading the enemy." *Id.* at 464 (citation omitted).

[87] This situation could be plausible if the AI had "learned" of the potential for misuse of the emblem or had been confused by data evasion and instead relied on a proxy characteristic it determined to be more effective. It is not unusual for machine learning to make conclusions based on non-traditional considerations. For example, AI has been shown to be able to determine the race of an individual solely by examining medical images (e.g., x-rays, CT scans) without reference to pictures showing the color of their skin. Banerjee *et al.*, Reading Race: AI Recognises Patient's Racial Identity in Medical Images, arXiv>Computer Science>Computer Vision and Pattern Recognition, (July 21, 2021), https://arxiv.org/abs/2107.10356.

this linkage, do they commit perfidy if they start carrying the same messenger bag? From a strict elements analysis of Article 37 of Additional Protocol I, a case could be made that perfidy was committed—specifically, that the combatant carried the message bag with the intent to invite the adversary to believe that the combatant was protected under the LOAC.[88]

A possible counterargument, and one that will become increasingly relevant in the era of greater reliance on AI in targeting decisions data science-driven warfare, is that no perfidy occurred because no human refrained from attack. One can envision a situation in the not-too-distant-future where data is gathered directly by technological sensors and fed directly to the AI for processing and action with little to no human involvement. AI systems will be able to increasingly take significant actions, and much more quickly, without meaningful human input. Where the data goes directly to the machine learning algorithm and a decision is made without human interaction, can there be perfidy? An argument can be made that there was no perfidy under a strict elemental analysis because "inviting the confidence of the adversary" presupposes a human adversary. Such an argument, on the other hand, seemingly runs counter to the basic object and purpose of the corpus of the LOAC. The danger of this approach is demonstrated by the fact pattern in the Scenario chapter, where the AI system attacked a doctor due to the confusion generated by enemy usage of data evasion patterns that unwittingly simulated medical personnel. Reconsideration of how perfidy applies in these circumstances should be considered in the data- and sensor-driven environment.

Recent scholarship has identified gaps in the perfidy rules which create confusion and the potential for problematic outcomes.[89] These gaps will only continue to grow with the introduction of machines into the decision cycle. This technological development will hopefully spur the accompanying development of custom and legal frameworks that appropriately address these developments.

VI. Conclusion: Big Data is Different

As we enter the era of big data and machine learning, it is important for legal scholars to view this technological development as more than simply the next step in the cyber revolution. In one sense, scholars are beginning to appreciate this fact as they begin to examine the legal implications of artificial intelligence in the battlefield. On the other hand, the current focus on artificial intelligence overlooks the big data that powers the artificial intelligence. Another risk is conflating big data with cyber operations generally. Big data is not merely a subset of cyber; while big data is cyber-enabled, it is

[88] Additional Protocol I, *supra* note 70, art. 37. Article 37 provides: "[i]t is prohibited to kill, injure or capture an adversary by resort to perfidy. Acts inviting the confidence of an adversary to lead him to believe that he is entitled to, or is obliged to accord, protection under the rules of international law applicable to armed conflict, with intent to betray that confidence, shall constitute perfidy." *Id.* To make the case in this hypothetical, one must prove that (1) the act (carrying the same messenger bag as ICRC personnel) is undertaken to invite the adversary that the soldier is protected by the Laws of Armed Conflict, (2) the adversary refrains from attacking because its AI concludes that the soldier is protected as a member of the ICRC, and (3) the soldier intentionally betrays the adversary's confidence by engaging in operations resulting in specified harmful consequences.

[89] *See generally*, Watts, *supra* note 85 (tracing the development of the law and identifying gaps in the definition of perfidy); Gary P. Corn and Peter P. Pascucci, The Law of Armed Conflict Implications of Covered or Concealed Cyber Operations: Perfidy, Ruses, and the Principle of Passive Distinction, *in* THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 273 (Eric Talbot Jensen & Ronald T.P. Alcala, eds., 2019); John C. Dehn, *Permissible Perfidy? Analyzing the Colombian Hostage Rescue, the Capture of Rebel Leaders and the World's Reaction*, 6 J. OF INT'L CRIM. JUST. 627 (2008)

fundamentally distinct both in its functioning and application. The connections that machine learning draws from various pieces of data are central the data mining/analytics process:

> "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem. Big Data is notable not because of its size, but because of its relationality to other data. Due to efforts to mine and aggregate data, Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself."[90]

In other words, it's not about the size of the data or its use in cyber operations alone, but instead the connections that can be derived from linking many disparate parts of the data, with applications far beyond the cyber realm. Attacks on big data in the future will focus on this feature, as can be seen by the rise of the field of Adversarial Machine Learning. The advent of big data will require legal frameworks that address this underlying foundational dynamic.

Data poisoning is but one aspect of Adversarial Machine Learning and can be expected on the battlefield before long. The analysis provided in this chapter demonstrates the inadequacy of merely "copying and pasting" pre-existing legal frameworks to this tactic. Legal scholars will need to become versed in the basic technical functioning of these systems, much like it was necessary to understand the basic functioning of computer networking in order to develop legal frameworks for cyber operations such as that found in the Tallinn Manual. Further, thoughtful consideration of the applicability of existing legal doctrines to big data will be required. As was demonstrated in this chapter, concepts such as proportionality or perfidy do not neatly map onto data poisoning attacks. Instead, novel and thoughtful analyses are needed. Similarly, the time frame of the applicability of a *lex specialis* such as the LOAC will be impacted by the practical realities of the timeframes of data poisoning operations.

The purpose of this chapter was to identify and frame some of the legal issues that are raised, not to provide definitive answers to these questions. These preliminary thoughts will hopefully be a starting point to begin working out the finer details of the needed legal analysis.

[90] danah boyd, Kate Crawford, *Six Provocations for Big Data*, OSF Preprints (Jan.4, 2017), osf.io/nrjhn.